A Language Modeling Approach to Predicting Reading Difficulty

Kevyn Collins-Thompson **Jamie Callan**
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
4502 Newell Simon Hall
Pittsburgh, PA 15213-8213
{kct, callan}@cs.cmu.edu

Abstract

We demonstrate a new research approach to the problem of predicting the reading difficulty of a text passage, by recasting readability in terms of statistical language modeling. We derive a measure based on an extension of multinomial naïve Bayes classification that combines multiple language models to estimate the most likely grade level for a given passage. The resulting classifier is not specific to any particular subject and can be trained with relatively little labeled data. We perform predictions for individual Web pages in English and compare our performance to widely-used semantic variables from traditional readability measures. We show that with minimal changes, the classifier may be retrained for use with French Web documents. For both English and French, the classifier maintains consistently good correlation with labeled grade level (0.63 to 0.79) across all test sets. Some traditional semantic variables such as type-token ratio gave the best performance on commercial calibrated test passages, while our language modeling approach gave better accuracy for Web documents and very short passages (less than 10 words).

1 Introduction

In the course of constructing a search engine for students, we wanted a method for retrieving Web pages that were not only relevant to a student's query, but also well-matched to their reading ability. Widely-used traditional readability formulas such as Flesch-Kincaid usually perform poorly in this scenario. Such formulas make certain assumptions about the text: for example, that the sample has at least 100 words and uses well-defined sentences. Neither of these assumptions need be true for Web pages or other non-traditional documents. We seek a more robust technique for predicting reading difficulty that works well on a wide variety of document types.

To do this, we turn to simple techniques from statistical language modeling. Advances in this field in the past 20 years, along with greater access to training data, make the application of such techniques to readability quite timely. While traditional formulas are based on linear regression with two or three variables, statistical language models can capture more detailed patterns of individual word usage. As we show in our evaluation, this generally results in better accuracy for Web documents and very short passages (less than 10 words). Another benefit of a language modeling approach is that we obtain a probability distribution across all grade models, not just a single grade prediction.

Statistical models of text rely on training data, so in Section 2 we describe our Web training corpus and note some trends that are evident in word usage. Section 3 summarizes related work on readability, focusing on existing vocabulary-based measures that can be thought of as simplified language model techniques. Section 4 defines the modified multinomial naïve Bayes model. Section 5 describes our smoothing and feature selection techniques. Section 6 evaluates our model's generalization performance, accuracy on short passages, and sensitivity to the amount of training data. Sections 7 and 8 discuss the evaluation results and give our observations and conclusions.

2 Description of Web Corpus

First, we define the following standard terms when referring to word frequencies in a corpus. A *token* is defined as any word occurrence in the collection. A *type* refers to a specific word-string, and is counted only once no matter how many times the word token of that type occurs in the collection.

For training our model, we were aware of no significant collection of Web pages labeled by reading difficulty level, so we assembled our own corpus. There are numerous commercial reading comprehension tests available that have graded passages, but this would have reduced the emphasis we wanted on Web documents. Also, some commercial packages themselves use read-

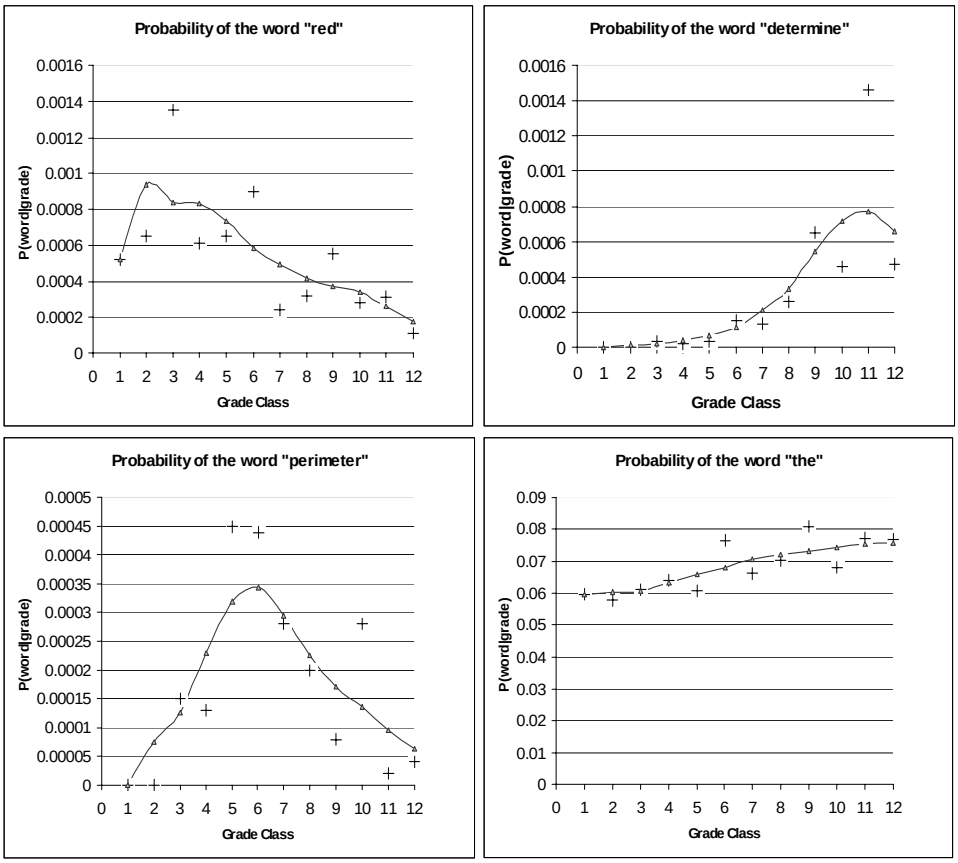

Figure 1. Examples of four different word usage trends across grades 1-12, as sampled from our 400K-token corpus of Web documents. Curves showing word frequency data smoothed across grades using kernel regression for the words (clockwise from top left): 'red', 'determine', 'the', and 'perimeter'.

ability measures when authoring the graded passages, making the data somewhat artificial and biased toward traditional semantic variables.

We gathered 550 English documents across 12 American grade levels, containing a total of 448,715 tokens and 17,928 types. The pages were drawn from a wide variety of subject areas: fiction, non-fiction, history, science, etc. We were interested in the accuracy available at individual grade levels, so we selected pages which had been assigned a specific grade level by the Web site author. For example, in some cases the assigned grade level was that of the classroom page where the document was acquired.

Before defining a classification model, we examined the corpus for trends in word frequency. One obvious pattern was that more difficult words were introduced at later grade levels. Earlier researchers (e.g. Chall, 1983, p. 63) have also observed that concrete words like 'red' become less likely in higher grades. Similarly, higher grade levels use more abstract words with increased frequency. We observed both types of behavior in our Web corpus. Figure 1 shows four words drawn from our corpus. Data from each of the 12 grades in the corpus are shown, ordered by ascending grade level. The solid line is a smoothed version of the word frequency data. The word 'red' does indeed show a steady decline in usage with grade level, while the probability of the word 'determine' increases. Other words like 'perimeter' attain maximum probability in a specific grade range, perhaps corresponding to the period in which these concepts are emphasized in the curriculum. The word 'the' is very common and varies less in frequency across grade levels.

Our main hypothesis in this work is that there are enough distinctive changes in word usage patterns between grade levels to give accurate predictions with simple language models, even when the subject domain of the documents is unrestricted.

3 Related Work

There is a significant body of work on readability that spans the last 70 years. A comprehensive summary of early readability work may be found in Chall (1958) and

Klare (1963). In 1985 a study by Mitchell (1985) reviewed 97 different reading comprehension tests, although few of these have gained wide use.

'Traditional' readability measures are those that rely on two main factors: the familiarity of semantic units (words or phrases) and the complexity of syntax. Measures that estimate semantic difficulty using a word list (as opposed to, say, number of syllables in a word) are termed 'vocabulary-based measures'.

Most similar to our work are the vocabulary-based measures, such as the Lexile measure (Stenner et al., 1988), the Revised Dale-Chall formula (Chall and Dale, 1995) and the Fry Short Passage measure (Fry, 1990). All of these use some type of word list to estimate semantic difficulty: Lexile (version 1.0) uses the Carroll-Davies-Richman corpus of 86,741 types (Carroll et al., 1971); Dale-Chall uses the Dale 3000 word list; and Fry's Short Passage Measure uses Dale & O'Rourke's 'The Living Word Vocabulary' of 43,000 types (Dale and O'Rourke, 1981). Each of these word lists may be thought of as a simplified language model. The model we present below may be thought of as a generalization of the vocabulary-based approach, in which we build multiple language models - in this study, one for each grade - that capture more fine-grained information about vocabulary usage.

To our knowledge, the only previous work which has considered a language modeling approach to readability is a preliminary study by Si and Callan (2001). Their work was limited to a single subject domain - science - and three broad ranges of difficulty. In contrast, our model is not specific to any subject and uses 12 individual grade models trained on a greatly expanded training set. While our model is also initially based on naïve Bayes, we do not treat each class as independent. Instead, we use a mixture of grade models, which greatly improves accuracy. We also do not include sentence length as a syntactic component. Si and Callan did not perform any analysis of feature selection methods so it is unclear whether their classifier was conflating topic prediction with difficulty prediction. In this paper we examine feature selection as well as our model's ability to generalize.

4 The Smoothed Unigram Model

Our statistical model is based on a variation of the multinomial naïve Bayes classifier, which we call the 'Smoothed Unigram' model. In text classification terms, each class is described by a language model corresponding to a predefined level of difficulty. For English Web pages, we trained 12 language models corresponding to the 12 American grade levels.

The language models we use are simple: they are based on unigrams and assume that the probability of a token is independent of the surrounding tokens, given the grade language model. A unigram language model is defined by a list of types (words) and their individual probabilities. Although this is a weak model, it can be trained from less data than more complex models, and turns out to give good accuracy for our problem.

4.1 Prediction with Multinomial Naïve Bayes

We define a *generative* model for a text passage T in which we assume T was created by a hypothetical author using the following algorithm:

1. Choose a grade language model G_i from some complete set of unigram models G according to a prior distribution $P(G_i)$. Each G_i has a multinomial distribution over a vocabulary V.

2. Choose a passage length L in tokens according to the distribution $P(L \mid G_i)$.

3. Assuming a 'bag of words' model for the passage, sample L tokens from G_i 's multinomial distribution based on the 'naïve' assumption that each token is independent of all other tokens in the passage, given the language model G_i.

The probability of T given model G_i is therefore:

$$P(T|G_i) = P(L|G_i) \cdot L! \prod_{w \in T} \frac{P(w|G_i)^{C(w)}}{C(w)!} \quad (1)$$

where $C(w)$ is the count of the type w in T.

Our goal is to find the most likely grade language model given the text T, or equivalently, the model G_i that maximizes $L(G_i|T) = \log P(G_i|T)$. We derive $L(G_i \mid T)$ from (1) via Bayes' Rule, which is:

$$P(G_i|T) = \frac{P(G_i)P(T|G_i)}{P(T)} \quad (2)$$

However, we first make two further assumptions:

1. All grades are equally likely *a priori*, and therefore $P(G_i) = 1/N_G$ where N_G is the number of grades.

2. The passage length probability $P(L|G_i)$ is independent of grade level.

Substituting (1) into (2), simplifying, and taking logarithms, we obtain:

$$L(G_i|T) = \sum_{w \in T} C(w) \log P(w|G_i) + \log Z \quad (3)$$

where $\log Z$ represents combined factors involving passage length and the uniform prior $P(G_i)$ which, according to our assumptions, do not influence the prediction outcome and may be ignored. The sum in (3) is easily computed: for each token in T, we simply look up its log probability in the language model of G_i and sum over all

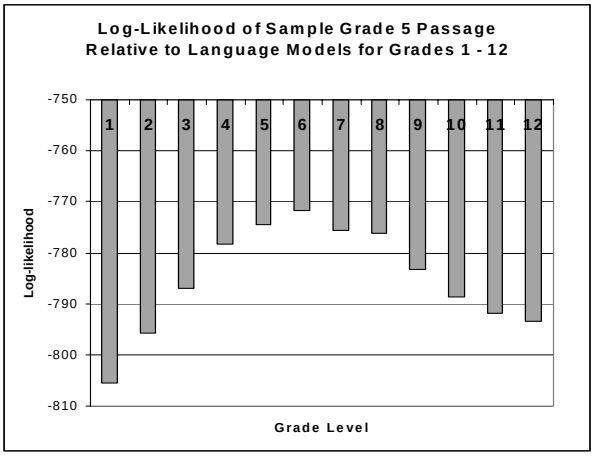

Figure 2. The log-likelihood of a typical 100-word Grade 5 passage relative to the language models for grades 1 to 12. The maximum log-likelihood in this example is achieved for the Grade 6 language model. Note the negative scale.

tokens to obtain the total likelihood of the passage given the grade. We do this for all language models, and select the one with maximum likelihood. An example of the set of log-likelihoods calculated across all 12 grade models, with a maximum point clearly evident, is shown in Figure 2.

5 Implementation

Given the above theoretical model, we describe two further aspects of our classification method: smoothing and feature selection.

5.1 Smoothing

We will likely see some types in test documents that are missing or rarely seen in our training documents. This is a well-known issue in language model applications, and it is standard to compensate for this sparseness by *smoothing* the frequencies in the trained models. To do this, we adjust our type probability estimates by shifting part of the model's probability mass from observed types to unseen and rare types.

We first apply smoothing to each grade's language model individually. We use a technique called Simple Good-Turing smoothing, which is a popular method for natural language applications. We omit the details here, which are available in Gale and Sampson (1995).

Next, we apply smoothing *across* grade language models. This is a departure from standard text classification methods, which treat the classes as independent. For reading difficulty, however, we hypothesize that nearby grade models are in fact highly related, so that even if a type is unobserved in one grade's training data, we can estimate its probability in the model by interpolating estimates from nearby grade models.

For example, suppose we wish to estimate $P(w|G)$ for a type w in a grade model G. If the type w occurs in at least one grade language model, we can perform regression with a Gaussian kernel (Hastie et al., 2001, p. 165) across all grade models to obtain a smoothed value for $P(w|G)$. With training, we found the optimal kernel width to be 2.5 grade levels. If w does not occur in *any* grade model (an 'out-of-vocabulary' type) we can back off to a traditional semantic variable. In this study, we used an estimate which is a function of type length:

$$\log P(w|G_i) \approx C + \frac{|w|}{D} \cdot (i - |w|)$$

where w is a type, i is a grade index between 1 and 12, $|w|$ is w's length in characters, and C = -13, D = 10 based on statistics from the Web corpus.

5.2 Feature Selection

Feature selection is an important step in text classification: it can lessen the computational burden by reducing the number of features and increase accuracy by removing 'noise' words having low predictive power.

The first feature selection step for many text classifiers is to remove the most frequent types ('stopwords'). This must be considered carefully for our problem: at lower grade levels, stopwords make up the majority of token occurrences and removing them may introduce bias. We therefore do not remove stopwords.

Another common step is to remove low-frequency types – typically those that occur less than 2 to 5 times in a model's training data. Because we smooth across grade models, we perform a modified version of this step, removing from all models any types occurring less than 3 times in the entire corpus.

Unlike the usual text classification scenario, we also wish to avoid some types that are highly grade-specific. For example, a type that is very frequent in the grade 3 model but that never occurs in any other model seems more likely to be site-specific noise than a genuine vocabulary item. We therefore remove any types occurring in less than 3 grade models, no matter how high their frequency. Further study is needed to explore ways to avoid over-fitting the classifier while reducing the expense of removing possibly useful features.

We investigated scoring each remaining type based on its estimated ability to predict (positively or negatively) a particular grade. We used a form of *Log-Odds Ratio*, which has been shown to give superior performance for multinomial naïve Bayes classifiers (Mladenic and Grobelnik, 1998). Our modified Log-Odds measure computes the largest absolute change in log-likelihood between a given grade and all other grades.

We tried various thresholds for our Log-Odds measure and found that the highest accuracy was achieved by using all remaining features.

5.3 Implementation Specifics

We found that we could reduce prediction variance with two changes to the model. First, rather than choosing the single most likely grade language model, we calculate the average grade level of the top *N* results, weighted by the relative differences in likelihood (essentially the *expected class*). The tradeoff is a small bias toward the middle grades. All results reported here use this averaging method, with *N=2*.

Second, to account for vocabulary variation within longer documents, we partition the document text into passages of 100 tokens each. We then obtain a grade level prediction for each passage. This creates a distribution of grade levels across the document. Previous work (Stenner, 1996, also citing Squires et al., 1983 and Crawford et al., 1975) suggests that a comprehension rate of 75% for a text is a desirable target. We therefore choose the grade level that lies at the 75th-percentile of the distribution, interpolating if necessary, to obtain our final prediction.

6 Evaluation

State-of-the-art performance for this classification task is hard to estimate. The results from the most closely related previous work (Si and Callan, 2001) are not directly comparable to ours; among other factors, their task used a dataset trained on science curriculum descriptions, not text written at different levels of difficulty. There also appear to be few reliable studies of human-human interlabeler agreement. A very limited study by Gartin et al. (1994) gave a mean interlabeler standard deviation of 1.67 grade levels, but this study was limited to just 3 samples across 10 judges. Nevertheless, we believe that an objective element to readability assessment exists, and we state our main results in terms of correlation with difficulty level, so that at least a broad comparison with existing measures is possible.

Our evaluation looked at four aspects of the model. First, we measured how well the model trained on our Web corpus generalized to other, previously unseen, test data. Second, we looked at the effect of passage length on accuracy. Third, we estimated the effect of additional training data on the accuracy of the model. Finally, we looked at how well the model could be extended to a language other than English – in this study, we give results for French.

6.1 Overall Accuracy and Generalization Ability

We used two methods for assessing how well our classifier generalizes beyond the Web training data. First, we applied 10-fold cross-validation on the Web corpus (Kohavi 1995). This chooses ten random partitions for each grade's training data such that 90% is used for training and 10% held back as a test set. Second, we used two previously unseen test sets: a set of 228 leveled documents from Reading A-Z.com, spanning grade 1 through grade 6; and 17 stories from Diagnostic Reading Scales (DRS) spanning grades 1.4 through 5.5. The Reading A-Z files were converted from PDF files using optical character recognition; spelling errors were corrected but sentence boundary errors were left intact to simulate the kinds of problems encountered with Web documents. The DRS files were noise-free.

Because the Smoothed Unigram classifier only models semantic and not syntactic difficulty, we compared its accuracy to predictions based on three widely-used *semantic* difficulty variables as shown below. All prediction methods used a 100-token window size.

1. **UNK**: The fraction of 'unknown' tokens in the text, relative to the Dale 3000 word list. This is the semantic variable of the Revised Dale-Chall measure.
2. **TYPES**: The number of types (unique words) in a 100-token passage.
3. **MLF**: The mean log frequency of the passage relative to a large English corpus. This is approximately the semantic variable of the unnormalized Lexile (version 1.0) score. Because the Carroll-Davies-Richman corpus was not available to us, we used the written subset of the British National Corpus (Burnard, 1995) which has 921,074 types. (We converted these to the American equivalents.)

We also included a fourth predictor: the Flesch-Kincaid score (Kincaid et al. 1975), which is a linear combination of the text's average sentence length (in tokens), and the average number of syllables per token. This was included for illustration purposes only, to verify the effect of syntactic noise. The results of the evaluation are summarized in Table 1.

On the DRS test collection, the TYPES and Flesch-Kincaid predictors had the best correlation with labeled grade level (0.93). TYPES also obtained the best correlation (0.86) for the Reading A-Z documents. However, Reading A-Z documents were written to pre-established criteria which includes objective factors such as type/token ratio (Reading A-Z.com, 2003), so it is not surprising that the correlation is high. The Smoothed Unigram measure achieved consistently good correlation (0.63 – 0.67) on both DRS and Reading A-Z test sets.

	Files	Grade Range	**Smoothed Unigram**	**UNK**	**TYPES**	**MLF**	**FK**
DRS	17	1.4 - 5.5	0.67	0.72	**0.93**	0.50	**0.93**
Reading A-Z	228	1.0 - 6.0	0.63	0.78	**0.86**	0.49	0.30
Web (Gr. 1-6)	250	1.0 - 6.0	**0.64**	0.38	0.26	0.36	0.25
Web (Gr. 1-12)	550	1.0 - 12	**0.79**	0.63	0.38	0.47	0.47

Table 1. Correlations between predictors and grade level, for the English collections used in our study. All predictors were trained on the Web corpus, with the Web tests using 10-fold cross-validation.

Flesch-Kincaid performs much more poorly for the Reading A-Z data, probably because of the noisy sentence structure. In general, mean log frequency (MLF) performed worse than expected – the reasons for this require further study but may be due to the fact the BNC corpus may not be representative enough of vocabulary found at earlier grades.

For Web data, we examined two subsets of the corpus: grades 1– 6 and grades 1– 12. The correlation of all variables with difficulty dropped substantially for Web grades 1–6, except for Smoothed Unigram, which stayed at roughly the same level (0.64) and was the best performer. The next best variable was UNK (0.38). For the entire Web grades 1– 12 data set, the Smoothed Unigram measure again achieved the best correlation (0.79). The next best predictor was again UNK (0.63). On the Web corpus, the largest portions of Smoothed Unigram's accuracy gains were achieved in grades 4– 8.

Without cross-grade smoothing, correlation for Web document predictions fell significantly, to 0.46 and 0.68 for the grade 1-6 and 1-12 subsets respectively.

We measured the type coverage of the language models created from our Web training corpus, using the Web (via cross-validation) and Reading A-Z test sets. Type coverage tells us how often on average a type from a test passage is found in our statistical model. On the Reading A-Z test set (Grades 1 – 6), we observed a mean type coverage of 89.1%, with a standard deviation of 6.65%. The mean type coverage for the Web corpus was 91.69%, with a standard deviation of 5.86%. These figures suggest that the 17,928 types in the training set are sufficient to give enough coverage of the test data that we only need to back off outside the language model-based estimates for an average of 8-10 tokens in any 100-token passage.

6.2 Effect of Passage Length on Accuracy

Most readability formulas become unreliable for passages of less than 100 tokens (Fry 1990). With Web applications, it is not uncommon for samples to contain as few as 10 tokens or less. For example, educational Web sites often segment a story or lesson into a series of image pages, with the only relevant page content being a caption. Short passages also arise for tasks such as estimating the reading difficulty of page titles, user queries, or questionnaire items. Our hypothesis was that the Smoothed Unigram model, having more fine-grained models of word usage, would be less sensitive to passage length and give superior accuracy for very short passages, compared to traditional semantic statistics.

In the extreme case, consider two single-word 'passages': 'bunny' and 'bulkheads'. Both words have two syllables and both occur 5 times in the Carroll-Davies-Richman corpus. A variable such as mean log frequency would assign identical difficulty to both of these passages, while our model would clearly distinguish them according to each word's grade usage.

To test this hypothesis, we formed passages of length L by sampling L consecutive tokens from near the center of each Reading A-Z test document. We compared the RMS error of the Smoothed Unigram prediction on these passages to that obtained from the UNK semantic variable. We computed different predictions for both methods by varying the passage length L from 3 tokens to 100 tokens.

The results are shown in Figure 3. Accuracy for the two methods was comparable for passages longer than about 50 tokens, but Smoothed Unigram obtained statistically significant improvements at the 0.05 level for 4, 5, 6, 7, and 8-word passages. In those cases, the prediction is accurate enough that very short passages may be reliably classified into low, medium, and high levels of difficulty.

6.3 Effect of Training Set Size on Accuracy

We derived the learning curve of our classifier as a function of the mean model training set size in tokens. The lowest mean RMS error of 1.92 was achieved at the maximum training set size threshold of 32,000 tokens per grade model. We fit a monotonically decreasing power-law function to the data points (Duda et al. 2001, p. 492). This gave extrapolated estimates for mean RMS error of about 1.79 at 64,000 tokens per model, 1.71 at 128,000 tokens per model, and 1.50 at 1,000,000 tokens per model.

While doubling the current mean training set size to 64,000 tokens per model would give a useful reduction in RMS error (about 6.7%), each further reduction of

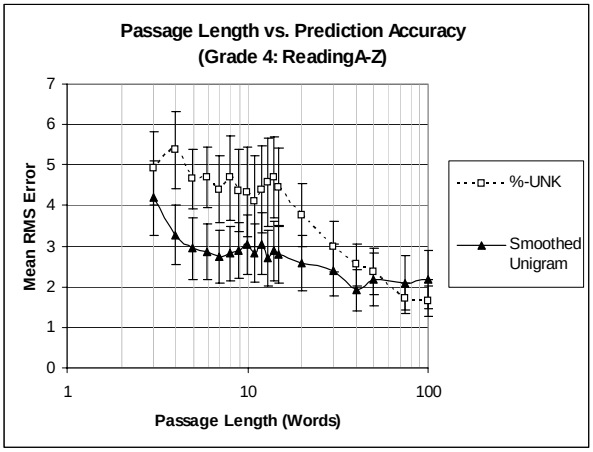

Figure 3. The effect of passage size on RMS prediction error for Grade 4 documents, comparing Smoothed Unigram to the UNK semantic variable. Error bars show 95% confidence interval. The grey vertical lines mark logarithmic length.

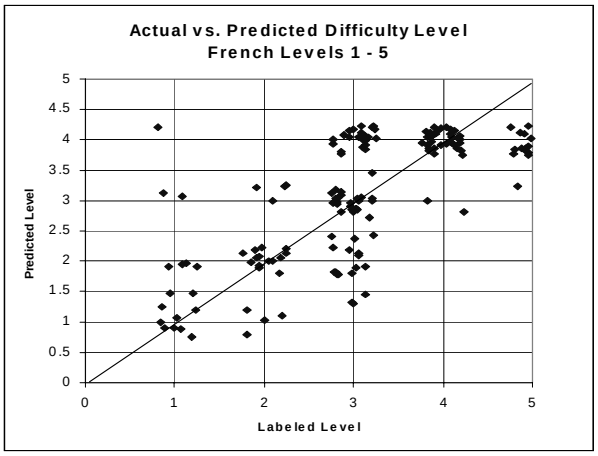

Figure 4. Actual vs. predicted difficulty label for documents from the French Web corpus. The data have been 'jittered' to show clusters more clearly. The diagonal line represents perfect prediction.

that magnitude would require a corresponding doubling of the training set size. This is the trade-off that must be considered between overall RMS accuracy and the cost of gathering labeled data.

6.4 Application to French Web Pages

To test the flexibility of our language model approach, we did a preliminary study for French reading difficulty prediction. We created a corpus of 189 French Web pages labeled at 5 levels of difficulty, containing a total of 394,410 tokens and 27,632 types (unstemmed).

The classification algorithm was identical to that used for English except for a minor change in the feature selection step. We found that, because of the inflected nature of French and the relatively small training set, we obtained better accuracy by normalizing types into 'type families' by using a simplified stemming algorithm that removed plurals, masculine/feminine endings, and basic verb endings.

A chart of the actual versus predicted difficulty label is shown in Figure 4. The classifier consistently underpredicts difficulty for the highest level, while somewhat over-predicting for the lowest level. This may be partly due to the bias toward central grades caused by averaging the top 2 predictions. More work on language-specific smoothing may also be needed. With 10-fold cross-validation, the French model obtained a mean correlation of 0.64 with labeled difficulty. For comparison, using the type/token ratio gave a mean correlation of 0.48. While further work and better training data are needed, the results seem promising given that only a few hours of effort were required to gather the French data and adjust the classifier's feature selection.

7 Discussion

While word difficulty is well-known to be an excellent predictor of reading difficulty (Chall & Edgar, 1995), it was not at all clear how effective our language model approach would be for predicting Web page reading difficulty. It was also unknown how much training data would be required to get good vocabulary coverage on Web data. Although retraining for other applications or domains may be desirable, two factors appear responsible for the fact that our classifier, trained on Web data, generalizes reasonably well to unseen test data from other sources.

First, smoothing across classes greatly reduces the training data required for individual grade models. By 'borrowing' word frequency data from nearby grades, the effective number of types for each grade model is multiplied by a factor of five or more. This helps explain the type coverage of about 90% on our test data.

Second, because we are interested in the relative likelihoods of grade levels, accurate relative type probabilities are more important than absolute probabilities. Indeed, trying to learn absolute type probabilities would be undesirable since it would fit the model too closely to whatever specific topics were in the training set. The important functions of relative likelihood appear to be general indicators such as the grade when a word is first introduced into usage, whether it generally increases or decreases with grade level, and whether it is most frequent in a particular grade range.

Further study is required to explore just how much this model of vocabulary usage can be generalized to other languages. Our results with French suggest that once we have normalized incoming types to accommo-

date the morphology of a language, the same core classifier approach may still be applicable, at least for some family of languages.

8 Conclusions

We have shown that reading difficulty can be estimated with a simple language modeling approach using a modified naïve Bayes classifier. The classifier's effectiveness is improved by explicitly modeling class relationships and smoothing frequency data across classes as well as within each class.

Our evaluation suggests that reasonably effective models can be trained with small amounts of easily-acquired data. While this data is less-rigorously graded, such material also greatly reduces the cost of creating a readability measure, making it easy to modify for specific tasks or populations.

As an example of retraining, we showed that the classifier obtained good correlation with difficulty for at least two languages, English and French, with the only algorithm difference being a change in the morphology handling during feature processing.

We also showed that the Smoothed Unigram method is robust for short passages and Web documents. Some traditional variables like type/token ratio gave excellent correlation with difficulty on commercial leveled passages, but the same statistics performed inconsistently on Web-based test sets. In contrast, the Smoothed Unigram method had good accuracy across all test sets.

The problem of reading difficulty prediction lies in an interesting region between classification and regression, with close connections to ordinal regression (MacCullagh, 1980) and discriminative ranking models (Crammer and Singer, 2001). While simple methods like modified naïve Bayes give reasonably good results, more sophisticated techniques may give more accurate predictions, especially at lower grades, where vocabulary progress is measured in months, not years.

Acknowledgements

This work was supported by NSF grant IIS-0096139 and Dept. of Education grant R305G03123. Any opinions, findings, conclusions, or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsors. We thank the anonymous reviewers for their comments and Luo Si for helpful discussions.

References

Burnard, L. (ed.) 1995. *The Users Reference Guide for the British National Corpus*. Oxford: Oxford University Computing Services.

Carroll, J. B., Davies, P., Richman, B. 1971. *Word Frequency Book.* Boston: Houghton Mifflin.

Chall, J.S. 1958. Readability: An appraisal of research and application. Bureau of Educational Research Monographs, No. 34. Columbus, OH: Ohio State Univ. Press.

Chall, J.S. 1983. *Stages of Reading Development*. McGraw-Hill.

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

Crammer, K. and Singer, Y. 2001. Pranking with ranking. *Proceedings of NIPS 2001*. 641-647.

Dale, E. and O'Rourke, J. 1981. *The Living Word Vocabulary.* Chicago, IL: World Book/Childcraft International.

Duda, R. O., Hart, P. E., and Stork, D. G. 2001. *Pattern Classification (Second Edition)*, Wiley, New York.

Fry, E. 1990. A readability formula for short passages. *J. of Reading*, May 1990, 594-597.

Gale, W., Sampson, G. 1995. Good-Turing frequency estimation without tears, *J. of Quant. Linguistics*, v. 2, 217-237.

Gartin, S., et al. 1994. W. Virginia Agriculture Teachers' Estimates of Magazine Article Readability. *J. Agr. Ed.* 35(1).

Hastie, T., Tibshirani, R., Friedman, J. 2001. *The Elements of Statistical Learning*. Springer-Verlag, New York.

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

Klare, G. R. 1963. *The Measurement of Readability.* Ames, IA. Iowa State University Press.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. of the 14th Int. Joint Conf. on Artificial Intelligence (IJCAI 1995)*. Montreal, Canada. 1137 - 1145.

MacCullagh, P. 1980. Regression models for ordinal data. *J. of the Royal Statistical Society B*, vol.42, 109-142.

Mitchell, J.V. 1985. *The Ninth Mental Measurements Yearbook*. Lincoln, Nebraska: Univ. of Nebraska Press.

Mladenic D., and Grobelnik, M. 1998. Feature selection for classification based on text hierarchy. *Working Notes of Learning from Text and the Web, CONALD-98*. Carnegie Mellon Univ., Pittsburgh, PA.

Reading A-Z.com 2003. Reading A-Z Leveling and Correlation Chart (HTML page). http://www.readinga-z.com/newfiles/correlate.html

Si, L. and Callan, J. 2001. A statistical model for scientific readability. *Proc. of CIKM 2001*. Atlanta, GA, 574-576.

Stenner, A. J., Horabin, I., Smith, D.R., and Smith, M. 1988. *The Lexile Framework.* Durham, NC: Metametrics.