A Machine Learning Approach to Reading Level Assessment

University of Washington CSE Technical Report 2006-06-06

Sarah E. Petersen*and Mari Ostendorf†

June 15, 2006

Abstract

Reading proficiency is a fundamental component of language competency. However, finding topical texts at an appropriate reading level for foreign and second language learners is a challenge for teachers. Existing measures of reading level are not well suited to this task, where students may know some difficult topic-related vocabulary items but not have the same level of sophistication in understanding complex sentence constructions. Recent work in this area has shown the benefit of using statistical language processing techniques. In this paper, we use support vector machines to combine features from statistical language models, traditional reading level measures, and other language processing tools to produce a better method of assessing reading level. We also discuss the performance of human annotators on this task.

1 Introduction

The U.S. educational system is faced with the challenging task of educating growing numbers of students for whom English is a second language [22]. In the 2001-2002 school year, Washington state had 72,215 students (7.2% of all students) in state programs for Limited English Proficient (LEP) students [2]. In the same year, one quarter of all public school students in California and one in seven students in Texas were classified as LEP [23]. Reading is a critical part of language and educational development, but finding appropriate reading material for LEP students is often difficult. To meet the needs of their students, bilingual education instructors seek out "high interest level" texts at low reading levels, e.g. texts at a first or second grade reading level that support the fifth grade science curriculum. Teachers also need to find material at a variety of

*Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195

†Department of Electrical Engineering, University of Washington, Seattle, WA 98195

levels, since students need different texts to read independently and with help from the teacher. Finding reading materials that fulfill these requirements is difficult and time-consuming, and teachers are often forced to rewrite texts themselves to suit the varied needs of their students.

Natural language processing technology is an ideal resource for automating the task of selecting appropriate reading material for bilingual students. Information retrieval systems successfully find topical materials and even answer complex queries in text databases and on the World Wide Web. However, an effective automated way to assess the reading level of the retrieved text is still needed. In this work, we develop a method of reading level assessment that uses support vector machines (SVMs) to combine features from statistical language models (LMs) and parse trees, with several traditional features used in reading level assessment. In preliminary work [18], we found that SVM-based detectors incorporating features from LMs and other sources outperformed LM-based detectors. In this paper, we present expanded results for the SVM detectors and describe experiments with human annotators to provide insights into the task difficulty.

The remainder of the paper is organized as follows. Section 2 describes related work on reading level assessment. Section 3 provides a brief introduction to support vector machines and statistical language models. Section 4 describes the corpora used in our work. In Section 5 we present our approach to the task, and in Section 6 we discuss experimental results. Section 7 discusses human annotations of our test data, to further explore the challenges inherent in this task. Section 8 provides a summary and description of future work.

2 Reading Level Assessment

This section highlights examples and features of some commonly used measures of reading level and discusses current research on the topic of reading level assessment.

Many traditional methods of reading level assessment focus on simple approximations of syntactic complexity such as sentence length. The widely-used Flesch-Kincaid Grade Level index is based on the average number of syllables per word and the average number of words per sentence in a passage of text [13] (as cited in [7]). Similarly, the Gunning Fog index is based on the average number of words per sentence and the percentage of words with three or more syllables [8]. These methods are quick and easy to calculate but have drawbacks: sentence length is not always an accurate measure of syntactic complexity, and syllable count does not necessarily indicate the difficulty of a word. A student may be familiar with a few complex words (e.g. dinosaur names) but unable to understand complex syntactic constructions.

Other measures of readability focus on semantics, which is usually approximated by word frequency with respect to a reference list or corpus. The Dale-Chall formula uses a combination of average sentence length and percentage of words not on a list of 3000 “easy” words [4]. The Lexile framework combines

measures of semantics, represented by word frequency counts, and syntax, represented by sentence length [20]. These measures are inadequate for our task; in many cases, teachers want materials with more difficult, topic-specific words but simple structure. Measures of reading level based on word lists do not capture this important information about structure. An additional drawback of these traditional approaches is that they use word lists that are updated manually. Our system is automatically trained, with the advantage that it can be customized for different levels or domains given only a small set of training documents.

In addition to the traditional reading level metrics, researchers at Carnegie Mellon University have applied probabilistic language modeling techniques to this task. Si and Callan [19] conducted preliminary work to classify science web pages using unigram models. More recently, Collins-Thompson and Callan manually collected a corpus of web pages ranked by grade level and observed that vocabulary words are not distributed evenly across grade levels. They developed a "smoothed unigram" classifier to better capture the variance in word usage across grade levels [7]. On web text, their classifier outperformed several other measures of semantic difficulty: the fraction of unknown words in the text, the mean log frequency of the text relative to a large corpus, and the Flesch-Kincaid measure. The traditional measures performed better on some commercial corpora, but these corpora were calibrated using similar measures, so this is not a fair comparison. More importantly, the smoothed unigram measure worked better on the web corpus, especially on short passages. The smoothed unigram classifier is also more generalizable, since it can be trained on any collection of data. Traditional measures such as Dale-Chall and Lexile are based on static word lists.

Although the smoothed unigram classifier outperforms other vocabulary-based semantic measures, it does not capture syntactic information. We believe that higher order n-gram models can achieve better performance by capturing both semantic and syntactic information. Additionally, an automatic parser can provide additional syntactic features. This is particularly important for the tasks we are interested in, when the vocabulary (i.e. topic) and grade level are not necessarily well-matched.

3 Classifiers

Support vector machines and statistical language models are topics which are well-documented in the literature. Hence, in this section, we provide only a very brief overview of both techniques.

3.1 Support Vector Machines

Support vector machines are a machine learning technique used in a variety of text classification problems. SVMs are based on the principle of structural risk minimization. Viewing the data as points in a high-dimensional feature space,

Table 1: Distribution of articles and words in the Weekly Reader corpus.

Grade Level	Number of Articles	Number of Words	Article Length (Words)	
			Mean	Std Dev
2	351	71.5k	161.1	146.5
3	589	444k	151.4	174.6
4	766	927k	254.3	197.8
5	691	1M	314.4	264.4

the goal is to fit a hyperplane between the positive and negative examples so as to maximize the distance between the data points and the plane. SVMs were introduced by Vapnik [24] and were popularized in the area of text classification by Joachims [11].

For training SVMs, we used the SVM^{light} toolkit developed by Joachims [12]. Using development data, we selected the radial basis function kernel and tuned parameters using cross validation and grid search as described in [10].

3.2 Statistical Language Models

Statistical LMs predict the probability that a particular word sequence will occur. The most commonly used statistical language model is the n-gram model, which assumes that the word sequence is an $(n-1)$th order Markov process. For example, for the common trigram model where $n = 3$, the probability of sequence $w = w_1, \ldots, w_m$ is:

$$P(w) = P(w_1)P(w_2|w_1)\prod_{i=3}^{m} P(w_i|w_{i-1}, w_{i-2}). \quad (1)$$

The parameters of the model are typically estimated using a maximum likelihood estimate based on the observed frequency in a training corpus and then smoothed. We used the SRI Language Modeling Toolkit [21] for language model training, with modified Kneser-Ney smoothing [6].

Perplexity (PP) is an information-theoretic measure often used to assess language models:

$$PP = \exp\left(-\frac{1}{m}\log P(w)\right). \quad (2)$$

Low perplexity indicates a better match between the test data and the model, corresponding to a higher probability $P(w)$. Here, we will use perplexity as a feature in the SVM.

4 Corpora

Our detectors are trained and tested on a corpus obtained from Weekly Reader, an educational newspaper with versions targeted at different grade levels [26].

Table 2: Number of articles in the Weekly Reader corpus as divided into training, development and evaluation test sets.

Grade	Training	Dev	Eval
2	315	18	18
3	529	30	30
4	690	38	38
5	623	34	34

These data consist of short articles on a variety of non-fiction topics, including science, history, and current events. Our corpus consists of articles from the second, third, fourth, and fifth grade editions of the newspaper because these grade levels were available in electronic form. This corpus contains just under 2400 articles, distributed as shown in Table 1. This table also includes the mean and standard deviation of the article lengths (in words), although article length was not used as a feature for our detectors. We want our detectors to perform well on a variety of texts, not just Weekly Reader articles, and article length is closely tied to this particular domain. In general, it is intuitive that lower grade levels often have shorter texts, but we would like to be able to classify short and long texts of all levels without assuming that short length is always an indicator of low reading level.

We divide the Weekly Reader corpus into separate training, development, and test sets, as shown in Table 2. The development data is used as a test set for tuning parameters, and the results presented in Section 6 are based on the evaluation test set. The development and evaluation test sets are the same size, and each consist of approximately 5% of the data for each grade level.

Additionally, we have two smaller corpora consisting of articles for adults and corresponding simplified versions for children or other language learners. Barzilay and Elhadad [1] have allowed us to use their corpus from Encyclopedia Britannica, which contains articles from the full version of the encyclopedia and corresponding articles from Britannica Elementary, a new version targeted at children. The Western/Pacific Literacy Network's [27] web site has an archive of CNN news stories and abridged versions that we have also received permission to use. Although these corpora do not provide an explicit grade-level ranking for each article, the adult and child/language-learner versions allow us to learn models that distinguish broad reading level categories. We can use these models to score articles from the Weekly Reader corpus or other sources to provide additional features for detection.

We use one other corpus in training, consisting of Associated Press newswire data from the TIPSTER corpus [9]. These are newspaper articles on a variety of topics; we selected this corpus as an example of text at an adult reading level in the same non-fiction/news domain as the Weekly Reader corpus. We use this corpus as "negative training data" to improve the accuracy of our detectors on text outside the Weekly Reader corpus. We will elaborate on this strategy in section 5. Table 3 shows the sizes of the supplemental corpora.

Table 3: Distribution of articles and words in supplemental training corpora.

Corpus	Num Articles	Num Words
Britannica	115	277k
Britannica Elementary	115	74k
CNN	111	51k
CNN Abridged	111	37k
TIPSTER Newswire	979	420k

Finally, for tests related to the generalizability of the approach, i.e. using data outside the Weekly Reader corpus, we downloaded 30 randomly selected newspaper articles from the "Kidspost" edition of The Washington Post [25]. We do not know the specific grade level of each article, only that "Kidspost" is intended for grades 3-8. We also downloaded 30 randomly chosen articles from the standard edition of The Washington Post.

5 Approach

In practice, a teacher is likely to be looking for texts at a particular level rather than classifying a group of articles into a variety of categories. A typical scenario is a teacher or student searching the Web (or other large collection of documents) for articles on a particular topic at a particular grade level. We would like to be able to filter articles by level just as search engines currently filter by topic. To address this scenario, we construct one detector per category which decides whether a document belongs in that category or not. In other words, the problem is posed as a detection task, rather than constructing a multi-class classifier that scores documents in terms of the specific classes for which there is labeled training data.

Existing reading level measures are inadequate for our intended task due to their reliance on vocabulary lists and/or a superficial representation of syntax. Our approach uses n-gram language models as a low-cost automatic approximation of both syntactic and semantic analysis, since syntactic dependencies are local in a large percentage of cases. Statistical LMs are used successfully in this way in other areas of language processing such as speech recognition and machine translation. We also use a standard statistical parser [5] to provide syntactic analysis. These and other features, described further in the next sections, are combined in a SVM framework to build grade level detectors. The unit of classification in this work is a single article.

5.1 Detector Features

Our SVM detectors for reading level use the following features:

- Traditional features:
  - Average sentence length

  - Average number of syllables per word
  - Flesch-Kincaid score
- 6 out-of-vocabulary (OOV) rate scores.
- Parse features (per sentence):
  - Average parse tree height
  - Average number of noun phrases
  - Average number of verb phrases
  - Average number of "SBAR"s.[1]
- 12 language model perplexity scores

The OOV scores are relative to the most common 100, 200 and 500 words in the lowest grade level (grade 2) in the training data. For each article, we calculated the percentage of a) all word instances (tokens) and b) all unique words (types) not on these lists, resulting in three token OOV rate features and three type OOV rate features per article.

The parse features are generated using the Charniak parser [5] trained on the standard Wall Street Journal Treebank corpus. We chose to use this standard data set as we do not have any domain-specific treebank data for training a parser. Although clearly there is a difference between news text for adults and articles intended for children, we presume that the parser can handle the simpler children's text and require it to handle adult text in the negative examples. Inspection of some of the resulting parses in the Weekly Reader corpus showed good accuracy.

Language model perplexity scores are used as indicators of semantic and syntactic similiarity to each of several reference corpora. If we had unlimited training data, these reference corpora would consist of additional Weekly Reader articles for each grade level. However, our corpus is limited and preliminary experiments in which the training data was split for LM and SVM training were unsuccessful due to the small size of the resulting data sets. Thus we made use of the Britannica and CNN corpora described in Section 4 to train trigram, bigram, and unigram models on each corpus of "child" text and "adult" text. This resulted in 12 LM perplexity features per article based on models of three n-gram orders trained on each of Britannica (adult), Britannica Elementary, CNN (adult) and CNN abridged text. Although these corpora do not map directly to Weekly Reader grade levels, they do represent broad differences in reading level and provide informative features for our detectors.

5.2 Feature Selection for N-gram Language Modeling

Feature selection is a common part of classifier design for many classification problems; however, there are mixed results in the literature on feature selection for text classification tasks. In Collins-Thompson and Callan's work [7] on readability assessment, LM smoothing techniques are more effective than other forms

[1] SBAR is defined in the Penn Treebank tag set as a "clause introduced by a (possibly empty) subordinating conjunction." It is an indicator of sentence complexity.

of explicit feature selection. However, feature selection proves to be important in other text classification work, including Lee and Myaeng's [14] genre detection work, which is similar to the reading level detection task. The approach used here was to use both feature selection and smoothing. Feature selection determines which words will be used as is vs. replaced by generic part-of-speech (POS) tokens. An n-gram language model is used to characterize the resulting mixed word-POS sequence, estimated using standard smoothing techniques.

Given $P(c|w)$, the probability of class c given word w, estimated empirically from the training set, we sorted words based on their information gain (IG) [28]. Information gain measures the difference in entropy when w is and is not included as a feature:

$$\begin{aligned} IG(w) = & - \sum_{c \in C} P(c) \log P(c) \\ & + P(w) \sum_{c \in C} P(c|w) \log P(c|w) + P(\bar{w}) \sum_{c \in C} P(c|\bar{w}) \log P(c|\bar{w}), \end{aligned} \quad (3)$$

and it corresponds to the mutual information between the class and the binary indicator random variable for word w. The most discriminative words are selected as features by plotting the sorted IG values and keeping only those words above the "knee" in the curve, as determined by manual inspection of the graph. All other words that appear in the text are replaced by their POS tag, as labeled by a maximum entropy tagger [17]. The resulting vocabulary consisted of 276 words and 56 POS tags. The use of POS tags was motivated by our goal of representing syntax: the tags allow the model to represent patterns in the text at a higher level than that of individual words, using sequences of POS tags to capture rough syntactic information. Early development experiments confirmed that the use of POS tags was much more effective than using a single generic word label.

6 Experiments

6.1 Evaluation Criteria

Results are assessed using multiple criteria. For analyzing our binary detectors, we use Detection Error Tradeoff (DET) curves and precision/recall measures. Detection Error Tradeoff curves show the tradeoff between misses and false alarms for different threshold values for the detectors. "Misses" are positive examples of a class that are misclassified as negative examples; "false alarms" are negative examples misclassified as positive. DET curves have been used in other detection tasks in language processing, e.g. Martin et al. [16]. We use these curves to visualize the tradeoff between the two types of errors, and select the minimum cost operating point in order to get a threshold for precision and recall calculations.

The minimum cost operating point on the DET curve depends on the relative costs of misses and false alarms. It is conceivable that one type of error might

Table 4: Precision, recall and F-measure on the test set for SVM-based detectors.

Grade	Precision	Recall	F-measure
2	38%	61%	47
3	38%	87%	53
4	70%	60%	65
5	75%	79%	77

be more serious than the other; however, teachers that we consulted with did not have a clear consensus as to which should be weighted higher. Hence, for the purpose of this analysis we weighted the two types of errors equally. In this work, the minimum cost operating point is selected by averaging the percentages of misses and false alarms at each point and choosing the point with the lowest average, which may not lie on the convex hull of the DET curve.

Different operating points on the DET curve correspond to different tradeoffs of precision and recall, where precision indicates the percentage of detected documents that match the target grade level, and recall indicates the percentage of the total number of target documents in the data set that are retrieved. Precision and recall are intuitively meaningful measures for this application, which is similar to information retrieval. However, because of the possibility of trading off one measure for gains in the other, it is sometimes difficult to compare systems using precision and recall, so the F-measure ($F = 2PR/(P + R)$) is often used to give a single system performance figure. Unless otherwise noted, precision, recall and F-measures reported are associated with the minimum cost operating point.

For comparison to other methods, e.g. Flesch-Kincaid and Lexile, which are not binary detectors, we consider the percentage of articles which are misclassified by more than one grade level. For our binary detection paradigm, this means that errors correspond to cases where articles at the target level are classified as more than one grade-level off (e.g. 2 classified as 4) by one or more detectors. Again, errors are calculated based on the minimum cost operating point.

6.2 Experiments with Weekly Reader Corpus

In this section, we present results for the SVM-based detectors trained and tested on the Weekly Reader corpus. Figures 1 and 2 show DET curves for detector performance on the development set and test set, respectively. The grade 2 and 5 detectors have the best performance, probably because in these cases the decision effectively involves only one boundary, whereas there are two boundaries (higher vs. lower) for the cases of grades 3 and 4. Using threshold values selected based on minimum cost on the development set, indicated by large dots on the plot, we calculated precision and recall on the test set, shown in Table 4. The grade 3 detector has high recall but relatively low precision;

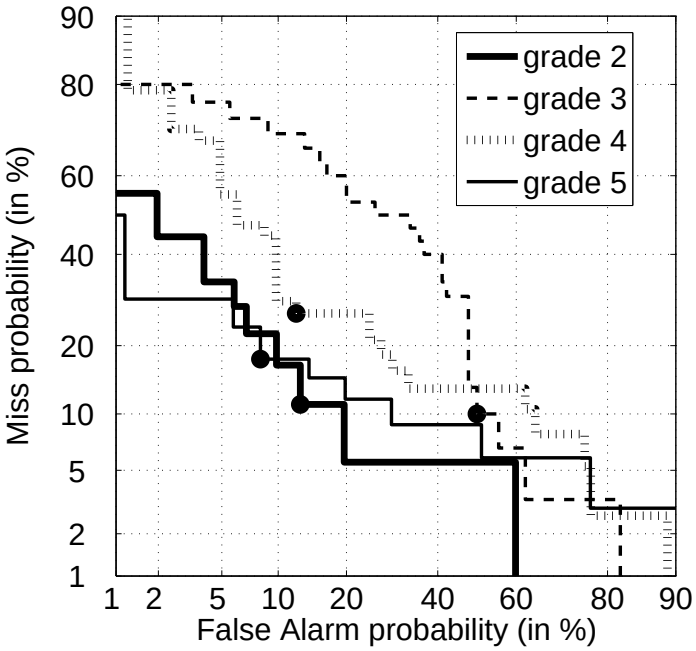

Figure 1: DET curves (development set) for SVM detectors with LM features.

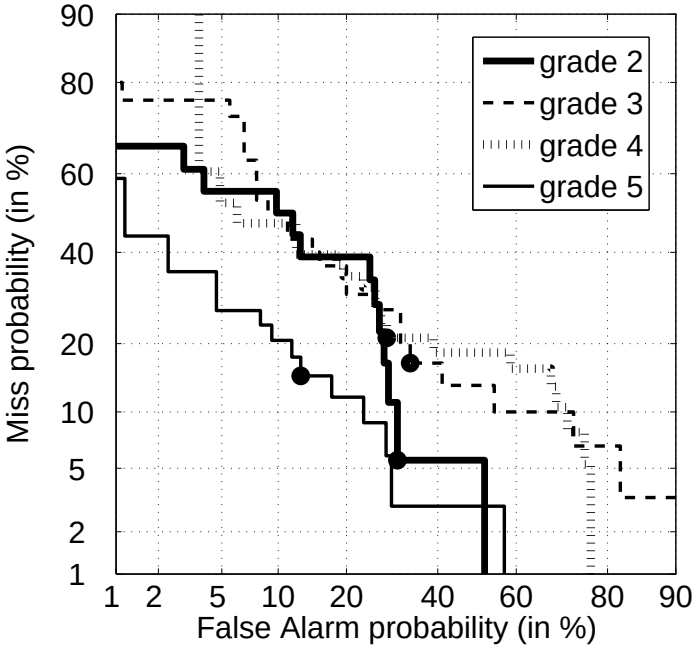

Figure 2: DET curves (test set) for SVM detectors with LM features.

Table 5: Number of Kidspost articles (out of 30) detected by each grade level detector for SVMs trained on the Weekly Reader (WR) data only vs. the WR data plus negative examples from TIPSTER newswire data. Articles that are not detected by any of the classifiers for grades 2-5 are counted under "Undetected".

Classifier Grade	**Classifier Training**	
	WR only	**WR + Newswire**
2	0	0
3	4	2
4	11	10
5	21	12
Undetected	0	12

the grade 4 detector does better on precision and reasonably well on recall. Since the minimum cost operating points do not correspond to the equal error rate (i.e. equal percentage of misses and false alarms) there is variation in the precision-recall tradeoff for the different grade level detectors, so F-measures are also given. For operating points chosen on the convex hull of the DET curves, the equal error rate ranges from 12-25% for the different grade levels.

We investigated the contribution of the four feature types to the overall performance of the SVM detector: traditional, OOV, parse, and LM features. We found that no feature type stood out clearly as most important, and performance was degraded when any particular category of features was removed. One trend we noted is that the traditional features help most in the higher grades (4, 5), i.e. performance drops most there when they are removed. The LM features are also more important for the higher grades; grade 2 actually improves a small amount without LM features. Performance generally drops when the OOV features are removed, and precision/recall tradeoff changes in favor of precision for grades 3, 4 and 5. The parse features are somewhat useful, particularly for grades 3 and 4, but they have less effect than other features.

6.3 Generalization Experiments

To assess the performance of the system on new data, the detectors were used with data downloaded from the "Kidspost" and standard editions of the Washington Post, as described in Section 4. In addition, since the detectors had only been trained on data reflecting reading levels 2-5, we trained new versions of the SVMs with TIPSTER newswire data included as additional negative training examples for each grade level. These negative training data were used to reduce the number of false positives for higher-level articles, particularly in the case of the grade 5 detector. It also leads to more realistic performance for the grade 5 detector on the lower-level articles, since the grade 5 detector now has the potential to reject articles as being at a higher as well as a lower level.

Table 5 includes detection results for the Kidspost articles for both the orig-

Table 6: Precision, recall and F-measure on test set. SVM detectors were trained on the Weekly Reader (WR) data (covering only grades 2-5) plus negative examples from TIPSTER newswire data.

Grade	**WR + Newswire**		
	Precision	**Recall**	**F-measure**
2	38%	78%	51
3	56%	83%	67
4	66%	55%	60
5	74%	68%	71

inal SVM detectors and the new version with augmented training data. Since the Kidspost data is targeted for the 3-8 grade range, one would expect that some of these articles would be above the grade 5 level. As we expected, the model trained only with Weekly Reader data detects a much larger number of articles at grade 5. This model also fails to leave any article unclassified, which is unrealistic since the Kidspost grade range is larger than the range of Weekly Reader data. The detectors trained on Weekly Reader and newswire data detect a more reasonable percentage of articles at grade 5 and leave 12 articles unclassified.

The benefit of the augmented training is particularly notable with the 30 articles from the standard edition of The Washington Post. All 30 of these articles were classified positively by the original grade 5 detector; the detector trained with newswire data as additional negative training data only positively classified 3 of these higher-level articles.

Adding newswire data as additional training data does change the performance of the new detectors on the original Weekly Reader corpus, as shown in Table 6. Recall improves for the grade 2 detector, as does precision for grade 3. There are small losses in both precision and recall in other cases, leading to more balanced F-measures for the detectors for grades 3-5. Figure 3 shows the differences in F-measures for the original SVM detectors trained on Weekly Reader data alone vs. the SVM detectors trained on Weekly Reader plus TIPSTER newswire data. The F-measures for the lower two grades improve with the addition of newswire data. While the higher grades have slightly worse performance, we believe that the numbers are more representative of the real task, and the advantage in generalization performance on other data is of substantial real-world importance.

6.4 Comparison with Other Methods

We also compared error rates for both versions of our SVM detectors (with and without newswire data) with two traditional reading level measures, Flesch-Kincaid and Lexile. The Flesch-Kincaid Grade Level index is a commonly used measure of reading level based on the average number of syllables per word and average sentence length. The Flesch-Kincaid score for a document is intended

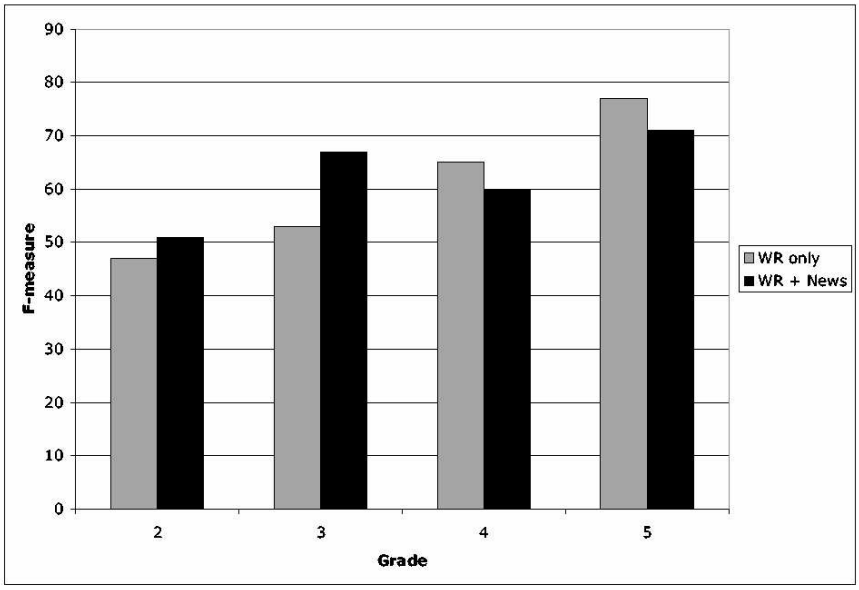

Figure 3: Comparison of F-measures for SVM detectors trained on the Weekly Reader (WR) data only, covering only grades 2-5, vs. the WR data plus negative examples from TIPSTER newswire data.

to directly correspond with its grade level. We chose the Lexile measure as an example of a reading level classifier based on word lists.[2] Lexile scores do not correlate directly to numeric grade levels. However, a mapping of ranges of Lexile scores to their corresponding grade levels is available on the Lexile web site [15]. Each SVM detector is independent of the detectors for the other grades, so it is possible for more than one detector to return "true" for a given article. In this case, we use a pessimistic measure and count errors even if another detector gave the correct answer. For example, if the grade 2 and grade 4 detectors both returned "true" for a grade 4 article, it counts as one error in this analysis.

For each of these classifiers, Table 7 shows the percentage of articles which are misclassified by more than one grade level. Flesch-Kincaid performs poorly, as expected since its only features are sentence length and average syllable count. Although this index is commonly used, perhaps due to its simplicity, it is not accurate enough for our intended application. Our SVM detector also outperforms the Lexile metric. Lexile is a more general measure while our detector is trained on this particular domain, so the better performance of our model is not entirely surprising. Importantly, however, our detector is easily tuned to any corpus of interest.

[2]Other classifiers such as Dale-Chall do not have automatic software available.

Table 7: Percentage of articles which are misclassified by more than one grade level by traditional and SVM classifiers.

Grade	**Errors**			
	Traditional		**SVM**	
	Flesch-Kincaid	**Lexile**	**WR**	**WR + Newswire**
2	78%	33%	6%	0%
3	67%	27%	3%	3%
4	74%	26%	13%	13%
5	59%	24%	21%	9%

The SVM detector trained on Weekly Reader data plus negative examples from TIPSTER newswire data performs the same as the original SVM detector for the middle grades, but does better on grades 2 and 5. This actually indicates that the negative training data improves the detector at all levels, since fewer misclassifications of grade 2 articles result from fewer mistakes by the grade 4 and 5 detectors, and similarly fewer grade 5 misclassifications result from better performance for the grade 2 and 3 detectors.

7 Human Annotation

One of the challenges in the area of reading level assessment is knowing the right answer. In our experiments, we take the grade level assigned to each article in the corpus by the writers and editors of Weekly Reader as the "gold standard." However, we were interested to see how difficult this kind of annotation task is for human experts and how well human annotators agreed with the labels given in the corpus. In our informal discussions with teachers, we have learned that experienced teachers feel that they are able to easily identify whether or not a text is the appropriate grade level for their students. We would like to know how consistent their impressions are with formal labels and assessment tools.

To investigate this issue, we conducted a study of the performance of human annotators on the Weekly Reader data. We hired four experts to annotate our test corpus. The first annotator (referred to as A in the following discussion) was an elementary school bilingual education teacher. The other three annotators were graduate students in fields relevant to reading and pedagogy: one student was earning a Master's in Teaching in elementary education (referred to as B), one was an English student with a particular interest in pedagogy (referred to as C), and one was a linguistics student (referred to as D). We provided the annotators with example articles of each grade level chosen randomly from the training data. Then we asked them to read each article in the test set (unlabeled and in random order) and mark which grade level(s) they thought were appropriate. In a small number of cases, the annotators did mark more than one grade level for a single article. We included all of these annotations in our analysis, since this is comparable to the way our SVM detectors work.

Table 8: Matrix of kappa values for inter-rater agreement of pairs of annotators A, B, C and D.

	B	**C**	**D**
A	0.41	0.40	0.28
B	-	0.54	0.26
C	-	-	0.34

Since our detectors are independent, a single article can have hits from more than one detector; likewise, an article can be classified as more than one grade level by the human annotators.

One way to evaluate human annotation performance is to see whether or not the annotators are consistent with each other, using the Cohen's kappa statistic. Cohen's kappa is a measure of inter-rater reliability, used to characterize the consistency of subjective annotations by human labelers in a variety of domains, including natural language processing tasks [3]. Cohen's kappa is calculated by comparing pairs of annotations from two labelers. For this task, the annotators could choose one or more labels per article so in our kappa calculations we consider, for each article, whether or not it is annotated positively for each grade level. This allows us to treat each annotator as a set of four "virtual detectors," which also allows for a more fair comparison with the SVM detectors. Table 8 shows Cohen's kappa values for each pair of annotators. Kappa values between 0.4 and 0.6 indicate moderate agreement. Good agreement results in a kappa above 0.6, which did not occur in this case. All pairs of annotators A, B and C show moderate agreement, while annotator D, the linguistics student, has low agreement with the other three annotators. We consider D an outlier and will focus the rest of our analysis on the three annotators with pedagogical backgrounds. It appears that a background in education, not just language, contributes substantially to annotation ability for this task.

Figure 4 shows precision and recall percentages for annotators A, B and C. Precision indicates the percentage of articles annotated at a particular grade level, including those that are annotated with multiple levels, that are considered to be associated with that grade level in the Weekly Reader categorization. Recall indicates what percentage of articles at a particular WR grade level were annotated with that grade level, regardless of other grade levels which might also have been chosen. Thus, annotating multiple grades per article penalizes precision but benefits recall. We observe two interesting trends in this chart. First, all of the numbers are less than 70%, and some are significantly lower. This is a difficult task, even for people with appropriate education and preparation for the task. Second, some grade levels seem consistently harder than others. For example, precision is low for grade 3 for all three annotators. This could in part be an artifact of this particular task; the human labelers knew that there were no articles higher than grade 5 or lower than grade 2, so precision was higher for grades 2 and 5. However, it is still clear that this is a challenging task with some inherent ambiguity.

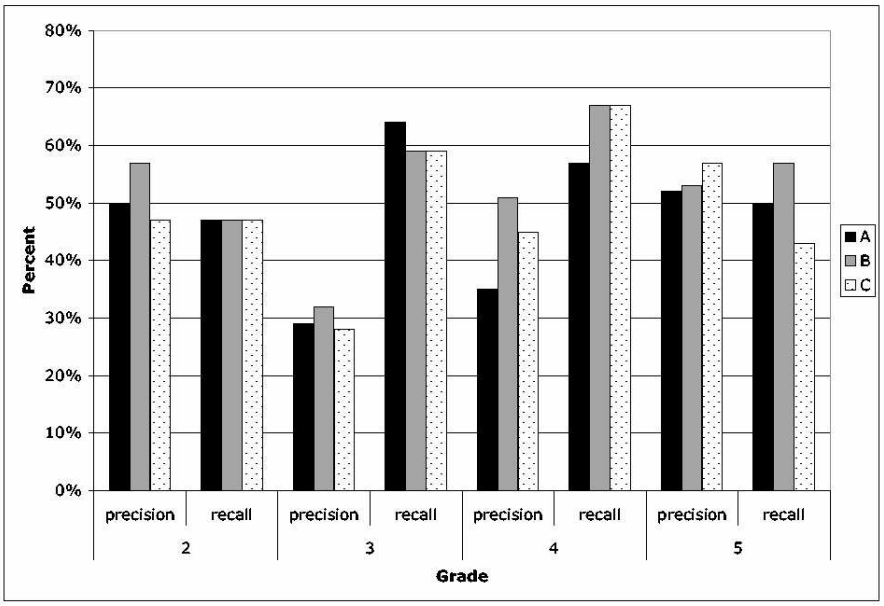

Figure 4: Precision and recall for annotators A, B, and C on the Weekly Reader test data set.

Next, we compare the performance of the annotators with the SVM detectors trained on Weekly Reader and TIPSTER newswire data. Table 9 shows the percentage of articles of each grade level which were misclassified by more than one grade by annotators A, B, and C and by the SVMs.[3] We note that annotator A has many more errors than the other annotators or the SVMs. This annotator marked an average of 1.5 grade levels per article, in comparison with an average of 1.2 for the other two annotators. More labels per article lead to both lower precision (on average) and more errors in this "off by more than one"

[3]The annotation experiments used a subset of about 80% of the original test set. The SVM results in this table are for this subset only and do not exactly correspond to the results in section 6.4.

Table 9: Percentage of articles which are misclassified by more than one grade level by annotators A, B and C and the SVM detectors.

Grade	**Errors**			
	A	**B**	**C**	**SVM**
2	59%	0%	12%	0%
3	9%	0%	0%	5%
4	10%	3%	3%	7%
5	23%	17%	30%	10%

scheme for this annotator. However, the other two annotators and the SVM detectors achieve roughly comparable performance, with the SVM outperforming the humans on grade 5. The SVM detectors are trained on this specific corpus, while the human annotators are likely to draw on personal experience as well as the sample articles and be better able to perform this task outside a specific corpus.

The results of this annotation study indicate that reading level assessment is a challenging problem for people with moderate experience and that the performance of the automatic system is comparable to that of human annotators. It may also be that there is an unavoidable bit of "noise" in the training and test data, associated with human variability for this task.

8 Conclusions and Future Work

We combine features from n-gram LMs, an automatic parser, and traditional methods of readability assessment in an SVM framework to classify texts based on reading level. We show that on a limited corpus, these detectors compare favorably to other existing methods and their performance is comparable to human annotators. These SVM detectors can be generalized to apply to newspaper text outside the inital domain with reasonable success. Adding higher-level negative training data improves generalization performance by reducing false positives without seriously degrading performance on the original test set. The SVM detectors are trainable, which makes it not surprising that they outperform general classifiers, but this is an important characteristic for tuning performance for the needs of particular groups (e.g., native language learners vs. second language learners) or even specific needs of individuals. Experiments with human annotators show that the task of reading level assessment is challenging for humans, particularly for those who have not training in elementary education. Further, we find that both the humans and the SVM detectors are better at detecting the lower grade levels. Future work includes working with teachers to evaluate the accuracy of our detectors on data from the World Wide Web and further refining the classifers using relevance feedback techniques.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0326276. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Thank you to Paul Heavenridge (Literacyworks), the Weekly Reader Corporation, Regina Barzilay (MIT) and Noemie Elhadad (Columbia University) for sharing their data and corpora.

References

[1] Barzilay, Regina and Nomie Elhadad. Sentence alignment for monolingual comparable corpora. In *Proc. of EMNLP*, pages 25–32, 2003.

[2] Bylsma, Pete, Lisa Ireland, and Helen Malagon. *Educating English Language Learners in Washington State*. Office of the Superintendent of Public Instruction, Olympia, WA, 2003.

[3] Carletta, Jean. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249-256, 1996.

[4] Chall, Jean S. and Edgar Dale. *Readability revisited: the new Dale-Chall readability formula*. Brookline Books, Cambridge, Mass., 1995.

[5] Charniak, Eugene. A maximum-entropy-inspired parser. In *Proc. of NAACL*, pages 132–139, 2000.

[6] Chen, Stanley and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359–393, 1999.

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

[8] Gunning, Robert. *The technique of clear writing*. McGraw-Hill, New York, 1952.

[9] Harman, Donna and Mark Liberman. *TIPSTER Complete.* Linguistic Data Consortium, catalog number LDC93T3A and ISBN: 1-58563-020-9. `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93T3A`, 1993.

[10] Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`, 2003. Accessed 11/2004.

[11] Joachims, Thorsten. Text categorization with support vector machines: learning with many relevant features. In *Proc. of the European Conference on Machine Learning*, pages 137–142, 1998a.

[12] Joachims, Thorsten. Making large-scale support vector machine learning practical. In *Advances in Kernel Methods: Support Vector Machines*. B. Schölkopf, C. Burges, A. Smola, eds. MIT Press, Cambridge, MA, 1998b.

[13] Kincaid, Jr., J.P., R.P. Fishburne, R.L. Rodgers, and B.S. Chisson. Derivation of new readability formulas for Navy enlisted personnel. Research Branch Report 8-75, U.S. Naval Air Station, Memphis, 1975.

[14] Lee, Yong-Bae and Sung Hyon Myaeng. Text genre classification with genre-revealing and subject-revealing features. In *Proc. of SIGIR*, pages 145–150, 2002.

[15] The Lexile framework for reading. `http://www.lexile.com`, 2005. Accessed April 15, 2005.

[16] Martin, A., G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. *Proc. of Eurospeech*, v. 4, pp. 1895-1898, 1997.

[17] Ratnaparkhi, Adwait. A maximum entropy part-of-speech tagger. In *Proc. of EMNLP*, pages 133–141, 1996.

[18] Schwarm, Sarah E. and Mari Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proc. of ACL*, pages 523–530, 2005.

[19] Si, Luo and Jamie Callan. A statistical model for scientific readability. In *Proc. of CIKM*, pages 574–576, 2001.

[20] Stenner, A. Jackson. Measuring reading comprehension with the Lexile framework. Presented at the Fourth North American Conference on Adolescent/Adult Literacy, 1996.

[21] Stolcke, Andreas. SRILM - an extensible language modeling toolkit. *Proc. ICSLP*, v. 2, pp. 901-904, 2002.

[22] U.S. Department of Education, National Center for Educational Statistics. The condition of education 2005. `http://nces.ed.gov/pubs2005/2005094.pdf`, Accessed November 17, 2005.

[23] U.S. Department of Education, National Center for Educational Statistics. NCES fast facts: Bilingual education/Limited English Proficient students. `http://nces.ed.gov/fastfacts/display.asp?id=96`, 2003. Accessed June 18, 2004.

[24] Vapnik, Vladimir. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

[25] The Washington Post. `http://www.washingtonpost.com`, 2005. Accessed April 20, 2005.

[26] Weekly Reader. `http://www.weeklyreader.com`, 2004. Accessed July, 2004.

[27] Western/Pacific Literacy Network / Literacyworks. CNN SF learning resources. `http://literacynet.org/cnnsf/`, 2004. Accessed June 15, 2004.

[28] Yang, Y. and J. Pedersen. A comparative study on feature selection in text categorization. *Proc. ICML*, pp. 412-420, 1997.