Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 23 (2009) 89–106
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

A machine learning approach to reading level assessment

Sarah E. Petersen [a,*], Mari Ostendorf [b]

[a] Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, United States
[b] Department of Electrical Engineering, University of Washington, Seattle, WA 98195, United States

Received 20 December 2006; received in revised form 22 April 2008; accepted 23 April 2008
Available online 7 May 2008

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 n-gram language models, parses, and traditional reading level measures to produce a better method of assessing reading level. We explore the use of negative training data to handle the problem of rejecting data from classes not seen in training, and compare the use of detection vs. regression models on this task. As in many language processing problems, we find substantial variability in human annotation of reading level, and explore ways that multiple human annotations can be used in comparative assessments of system performance.

Keywords: Reading level assessment; SVMs

1. Introduction

The US educational system is faced with the challenging task of educating growing numbers of students for whom English is a second language (US Department of Education, 2005). 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 (Bylsma et al., 2003). In the same year, one quarter of all public school students in California and one in seven students in Texas were classified as LEP (US Department of Education, 2003). 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 of teenagers who are poor readers face a similar problem.) Finding reading materials that fulfill these requirements is difficult and time-consuming, particularly when trying to meet

* Corresponding author.
E-mail addresses: sarahs@cs.washington.edu (S.E. Petersen), mo@ee.washington.edu (M. Ostendorf).

0885-2308/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2008.04.003

the needs of different students, and teachers are often forced to rewrite texts themselves to suit the varied needs of their students.

Natural language processing technology can help automate 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. Our strategy is to apply text classification techniques to this problem.

In preliminary work (Schwarm and Ostendorf, 2005), we developed a method of reading level assessment that uses support vector machines (SVMs) to combine features from n-gram language models (LMs) and parse trees, with several traditional features used in reading level assessment. 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, including:

- addressing the problem of generalizing the classifier to handle new data that may include other grade levels beyond those in the hand-labeled training data by introducing unlabeled negative training data (newswire text);
- investigating the degree to which syntactic features provide a benefit over traditional lexical features; and
- exploring the usefulness of a regression model as an alternative to the binary detection framework originally proposed, particularly in the context of limited annotated training data.

We also include experiments with human annotators to provide insights into the task difficulty and to present different methods of evaluating our detectors in comparison to existing approaches to reading level assessment.

In the sections to follow, we provide a bit more background on related research in reading level assessment, followed by a description of the data used, the details of the approach and the experiment results. Experimental results strengthen our earlier findings that SVMs outperform traditional methods for reading level assessment and findings in other work that syntactic features provide only a small benefit for this task. They also show that the detection model is a better match to this problem than a regression model, at least for the case where annotated training data is limited. In the work with human annotations, we find that the humans actually do less well than the SVM in labeling the target data, indicating that different groups may use different criteria for reading level assessment and supporting the use of machine learning as means of tuning the decision function to the needs of a particular group.

2. Reading level assessment

The process used by teachers to select appropriate reading material for their students is complicated and subjective, taking into account subject matter as well as characteristics of the text itself. For example, Fountas and Pinnell's well-known system of matching books to readers takes into account more than a dozen high-level characteristics, including vocabulary, grammatical structure of phrases and sentences, use of literary devices, illustrations, and layout on the page (Fountas and Pinnell, 1999). Automatic tools cannot capture this entire range of characteristics, but a variety of methods and formulae have been developed to calculate approximations of reading level based on characteristics which are easily measured.

Many traditional formulae for 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 (Kincaid et al, 1975) (as cited by Collins-Thompson and Callan (2005)). 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 (Gunning, 1952). 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. Also, 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 found on a list of 3000 "easy" words (Chall and Dale, 1995). The Lexile frame-

work combines measures of semantics, represented by word frequency counts (based on a 600 million word corpus), and syntax, represented by sentence length (Stenner, 1996). Again, these measures are inadequate for the task of finding materials with more difficult, topic-specific words but simple structure. Measures of reading level based on word lists do not capture this information about structure. An additional drawback of some of these traditional approaches, e.g., Dale–Chall, is that they use word lists that are updated manually.

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 (2001) 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 (Collins-Thompson and Callan, 2005). 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 it is arguably not a fair comparison. More importantly, the smoothed unigram measure worked better on the web corpus, especially on short passages.

Although the smoothed unigram classifier outperforms other vocabulary-based semantic measures, it does not capture syntactic information. Our conversations with bilingual education and ESL teachers suggested that this would be a useful additional feature; hence, we included parse features and part-of-speech tag sequence probabilities in our initial work (Schwarm and Ostendorf, 2005). However, we did not explicitly measure the contribution of these features. In subsequent work, Hellman et al. (2007) argue that grammatical features are more relevant for second language (L2) acquisition than for first language (L1) readers, since most grammar in the first language is acquired prior to the start of formal education, unlike in the L2 case. They explore grammatical constructions identified from textbooks for three ESL levels as features in reading level assessment. They find that lexical features alone outperform grammatical features, but that there is a performance gain from combining both (more for L2 than L1 conditions). The results we report here, analyzing the contributions of our lexical and syntactic features, are consistent with these findings.

3. Text classification

The task of reading level assessment can be viewed as a type of text classification, where text classification includes topic labeling, genre detection, author identification, etc. The basic machine learning techniques used typically generalize to all of these problems, though the features can be quite different.

Most relevant to reading level classification is the work on genre classification, because of the emphasis on style over topic. Early work on genre detection (Kessler et al., 1997) identified structural cues from tagged/parsed text, lexical cues such as dates and terms of address (e.g., Mrs.), character-level cues such as punctuation and capitalization, and derivative cues (e.g., ratios of the above features). Stamatatos et al. (2000) use discriminant analysis and find that the frequencies of the most common words in the text (function words) are the most important features, though punctuation is also helpful. Lee and Myaeng (2002) use an extension of an information retrieval statistic, term frequency-inverse document frequency (TF-IDF), to select genre-revealing terms. The goal is to find words that appear often in a particular genre and not others and are evenly distributed across topics within that genre.

The word-based features used by all these systems are based on the bag-of-words representation, i.e., words in the sequence are treated independently, which is probably the most popular representation. However, there are several text classification systems that use n-gram models to characterize word sequences, both as count-based features (Damashek, 1995, Huffman, 1995) in a vector space model and using class-dependent language models directly as the classifier (Yamron et al., 1999, Peng et al., 2003). Here, smoothing takes the place of other forms of feature selection commonly used in classification algorithms. While much of the work with n-grams has been on topic classification, which is very different from reading level detection, the work by Peng et al. (2003) shows success on a variety of tasks, including genre detection. Language model probabilities can also be used as features in another classifier, such as an SVM, which allows the use of multiple language

models. SVMs have been used with much success in text classification problems (Joachims, 1998), and we use them in this work for that reason.

4. Corpora

One challenge in this work was finding an appropriate corpus. A large collection of texts with reading levels labeled for our target audience of LEP students would be ideal; however, we are not aware of any such collection that exists in electronic form. Instead, we made use of an available set of texts aimed at child language learners with reading level indicated. This data is useful for demonstrating the impact of automatically trained detectors, which could later be trained with any annotated corpus.

The detectors are trained and tested on a corpus obtained from Weekly Reader, an educational newspaper with versions targeted at different grade levels (Weekly Reader, 2004). These data consist of short articles on a variety of non-fiction topics, including science, history, and current events. The 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. These US grade levels correspond approximately to ages 7–10. 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 the detectors. 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 (2003) provided 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. We also received permission to use the archive of CNN news stories and abridged versions at the Western/Pacific Literacy Network (2004) web site. Although these corpora do not provide an explicit grade-level ranking for each article, the adult and child/language-learner versions allow us to train language models that distinguish broad reading level categories. We use these language models to score articles from the Weekly Reader corpus and other sources to provide features for detection.

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

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

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

We use one other corpus in training, consisting of Associated Press newswire data from the TIPSTER corpus (Harman and Liberman, 1993). These are 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. While the reading level is not indicated and the level likely varies within the corpus, it is probably safe to assume that it is higher than grade 5. We use this corpus as "negative training data" to improve the accuracy of our detectors on text outside the Weekly Reader corpus, since otherwise the classifier cannot distinguish between grade 5 and higher levels. As the experiments will show, it is useful for this purpose, although less so for training a regression model where it may be more important to have specific grade-level annotations. Table 3 shows the sizes of the supplemental corpora.

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 The Washington Post (2005). 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

There are two different ways in which reading level assessment tools could be used. In the first case, we imagine a teacher who is looking for texts at a particular level for an individual student or class. In the second case, we want to classify a group of articles into a variety of categories, perhaps for inclusion in a database. The first case corresponds to a binary detection problem, and the second involves either n-way classification or regression. In this work, we focus primarily on the first case, in which a typical scenario is a teacher or student searching the Web (or other large collection of documents) for articles on a certain 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. However, we also include some experiments using regression for more direct comparison to other techniques and for comparison to the detection approach.

To address the detection scenario, we construct one detector per category which decides whether an article belongs in that category or not. To address the second scenario, we train a regression model and round the predicted continuous value to the nearest integer grade level. In both cases we use SVMs, because of their prior success in text classification problems (Joachims, 1998). For training SVMs, we used the SVM^{light} toolkit developed by Joachims (1999). Using development data, we selected the radial basis function kernel and tuned parameters using cross validation and grid search as described by Hsu et al. (2003).[1]

The features used in this work are the same as in Schwarm and Ostendorf (2005), but are described below for completeness, since one goal of this work is to assess the contribution of different types of features.

5.1. Detector features

The particular features used in the SVMs are motivated by prior work (i.e., lexical features, syllable count, and sentence length) and by the goal of developing an online rating tool (i.e., relative low cost). They include:

[1] The best parameter set varied depending on grade level; the average values are $C = 41$, gamma $= 2 \times 10^{-4}$, and 1015 support vectors. Grade 2 had significantly fewer support vectors (568), probably because there are no classes lower than 2.

- Purely lexical features:
  - 6 out-of-vocabulary (OOV) rate scores
  - Average number of syllables per word
- Features that provide some representation of syntax:
  - 12 language model scores
  - 4 parse features
- Other traditional features:
  - Average sentence length
  - Flesch-Kincaid score

All features except for the parse features are low cost, and the parser is quite efficient given its high quality. The parser is used to assess the maximum possible gain from an automatically generated parse. Further details on the specific implementation follow.

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. These sizes covered 56%, 65% and 77% of the tokens in the full grade 2 training set, where we aimed low to avoid including too many topic-related words. (As expected, coverage rates are reduced as grade level increases, with overall coverages of 48%, 55% and 64% for grade 5.) 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 number of syllables per word came from a 85k word dictionary based on the Pronlex dictionary available from the Linguistic Data Consortium, which was syllabified using Fisher's implementation of Kahn's theory (Fisher, 1996). Syllable counts for words not found in the dictionary were generated by the publicly-available Perl module Lingua::En::Hyphenate.

Language models were trained on four different data resources: Britannica (adult), Britannica Elementary, CNN (adult) and CNN abridged. For each data set, three language models were trained: trigram, bigram, and unigram models. Together, there were 12 language models, all based on the same vocabulary (described in the next section). Each language model was used to score the target article by computing its perplexity, and the 12 perplexity values were used as features. (Low perplexity using models trained with the adult levels and high perplexity on the elementary/abridged levels would indicate a high reading level.) 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, and by using different corpora we avoid overtraining on the Weekly Reader data. The different order n-grams are used to provide a variety of features to the SVM, since we do not know a priori which will work best for this task.

The parse features include per-sentence averages of parse tree height, noun phrase count, verb phrase count, and SBAR count.[2] Parses are generated using the Charniak parser (Charniak, 2000) trained on the standard Wall Street Journal Treebank corpus, chosen due to our interest in a scenario where text is selected from the web, assuming that most texts will be more news-like than conversational. While the lower level (children's) text may have different distributions of syntactic constructions than the newspaper text, we assume that the WSJ corpus at least covers the simple types of constructions typically observed in text written for primary grade levels. Inspection of some of the resulting parses in the Weekly Reader corpus showed satisfactory results.

We used the Flesch-Kincaid score as a feature since this traditional measure of reading level can easily be calculated automatically. It is included to ensure that machine learning does at least as well as the baseline of only the Flesch-Kincaid score. If not, we would suspect an overtraining problem.

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 work by Collins-Thompson

[2] 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.

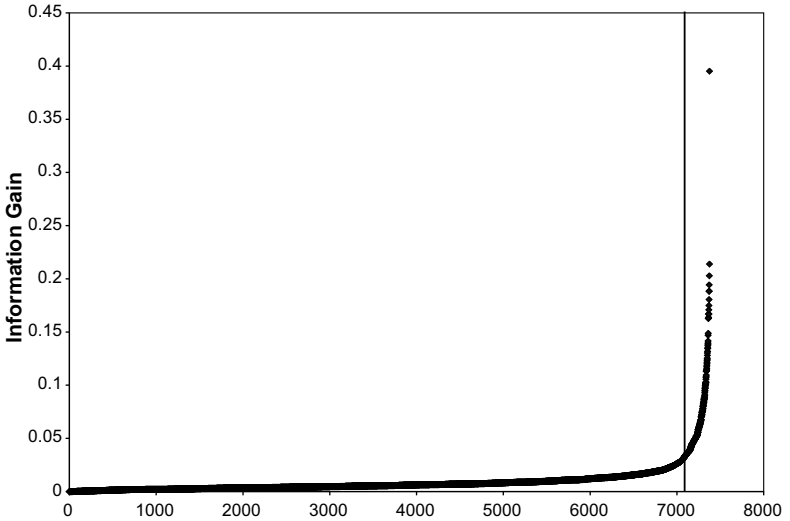


Fig. 1. Information gain of (ordered) words for feature selection.

and Callan (2005) on readability assessment, LM smoothing techniques are more effective than other forms of explicit feature selection. However, feature selection proves to be important in other text classification work, including genre detection (Lee and Myaeng, 2002), so we combine the two methods.

Our approach first uses feature selection to determine which words will be used as is vs. replaced by the part-of-speech (POS) tag associated with that word. N-gram language models with smoothing are then used to characterize the resulting mixed word/POS sequence, ignoring the original word labels of all tokens replaced with POS labels (thus different from a class language model) but not discarding the tokens entirely (hence a bit different from standard feature selection). Including smoothing on top of feature selection is useful for sequence modeling. Without smoothing, it is only practical to use the bag-of-words representation. With the sequence model, we can represent patterns in the text beyond the individual words, including salient word pairs but also a rough representation of syntax via the POS tags.[3] Early development experiments and our previous work using LM-only classifiers (Schwarm and Ostendorf, 2005) confirmed that the use of POS tags was much more effective than using a single generic word label, and that feature selection (mixed word-POS models) led to better performance than word-based models alone.

The specific approach to feature selection used information gain (IG) (Yang and Pedersen, 1997) to rank the most informative words for detecting reading level, based on training set class posteriors. Information gain measures the difference in entropy when word w is and is not included as a feature, and it corresponds to the mutual information between the class and the binary indicator variable for 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 (see Fig. 1). All other words that appear in the text are replaced by their POS tag, as labeled by a maximum entropy tagger (Ratnaparkhi, 1996). The resulting vocabulary consisted of 276 words and 56 POS tags. We used the SRI Language Modeling Toolkit (Stolcke, 2002) for language model training with a standard smoothing algorithm (modified Kneser-Ney smoothing (Chen and Goodman, 1999)). We experimented with using multiple thresholds for feature selection (which lead to different vocabularies and hence different language models), but there was no benefit in performance and some degradation was observed when using language models that differ in terms of vocabulary as well as the other factors.

[3] The use of POS n-grams as a stand-in for syntax is common place in text analysis for speech synthesis and in predicting punctuation for speech transcription.

6. Experiments

6.1. Evaluation criteria

The detectors are assessed primarily using 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. In the detection scenario, where an article can have multiple labels, precision considers the labels from all detectors, and recall considers only the labels from the detector for that grade level. Thus, positive detector results for multiple grades per article penalize precision but benefit recall. Due to the possibility of trading off one measure for gains in the other, the F-measure ($F = 2PR/(P + R)$) is often used to give a single system performance figure. Precision, recall and F-measures reported are associated with the minimum cost operating point on the detection-error tradeoff curve.

We also include some results where we compare systems based on the percentage of articles with labels that are off by more than one grade level, under the assumption that such errors are more problematic.

6.2. Baseline SVM results

Results for the baseline SVM detectors trained and tested on the Weekly Reader corpus (from Schwarm and Ostendorf (2005)) are shown in Table 4. The grade 3 detector has high recall but relatively low precision; the grade 4 detector does better on precision and reasonably well on recall. As illustrated in the DET curves shown in Fig. 2, the minimum cost operating points do not correspond to the equal error rate (i.e., equal percentage of false negatives and false positives), so there is variation in the precision–recall tradeoff for

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

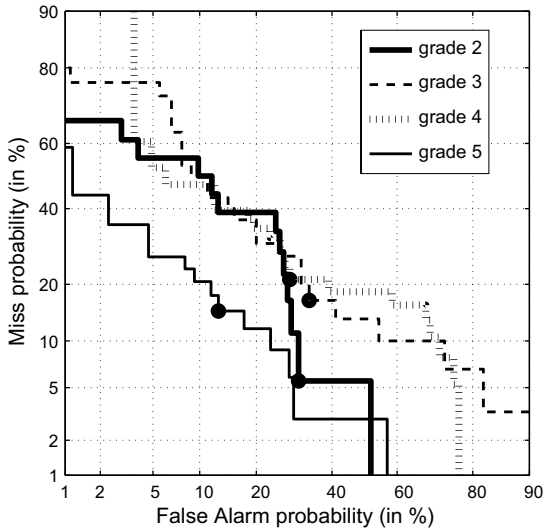


Fig. 2. DET curves (test set) for SVM detectors. The large dot on each curve indicates the minimum cost error point.

the different grade-level detectors. The fact that the grade 5 curve is significantly better than the others is in part due to the fact that there is no potential for labeling articles at a higher level, an issue we explore in the next section.

6.3. Generalization experiments

An important problem that this work seeks to solve is that of generalizing the classifier to new data, specifically discriminating the target grade levels (2–5) from other levels not seen in training. We address this problem in training by including negative training data (AP newswire) that is unmarked for grade level but known to be targeted at a much higher level than grade 5. The hope is that these negative training data will 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.

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 newspaper, as described in Section 4. Table 5 includes detection results for the Kidspost articles for both the original 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 and therefore not classified by our detectors. No information about the target grade range of the articles was provided to the SVM detectors. Both grade 2 detectors correctly rejected all 30 articles. As we expected, the detector trained only with Weekly Reader data detects a much larger number of articles at grade 5, failing to leave any article unclassified. The detectors trained on Weekly Reader plus 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, leaving the remaining 27 articles undetected.

Adding newswire data as additional training data does change the performance of the new detectors on the original Weekly Reader corpus. Fig. 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, these differences are not statistically significant. Furthermore, the percentage of articles where the detector is off by more than one level is the same or lower for the detector trained with the negative newswire data.

We also compared the performance of both versions of our SVM detectors (with and without newswire data) with a regression-based SVM classifier, trained using SVM^{light}'s regression mode, using the same features and training data. For the regression classifier, we use "9" as the target grade level for the newswire data. There is no grade level explicitly assigned to this corpus, but most news articles are targeted for a high school reading level. In classification with the regression model, we round the predicted value to the nearest level.

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

Classifier grade	Classifier training	
	WR only	WR+Newswire
2	0	0
3	4	2
4	11	10
5	21	12
Undetected	0	12

Articles that are not detected by any of the classifiers for grades 2–5 are counted under "Undetected".

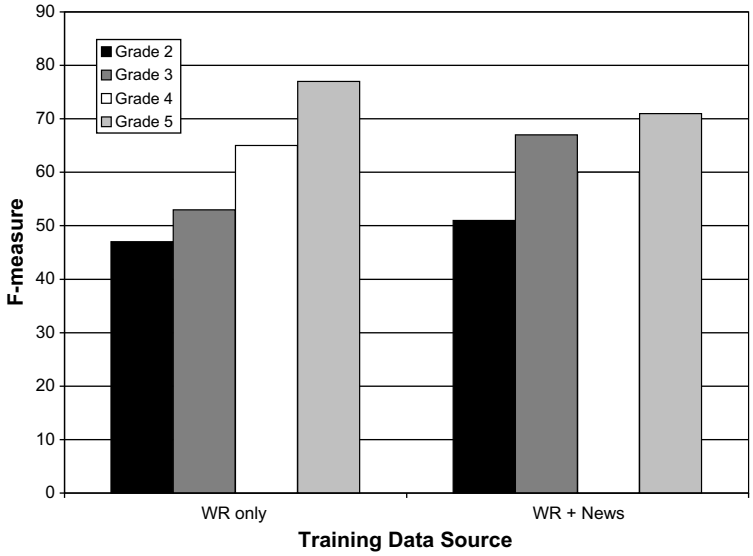


Fig. 3. Comparison of *F*-measures for SVM detectors trained on the Weekly Reader (WR) data only, covering only grades 2–5, vs. the WR data plus negative examples from TIPSTER newswire data.

Fig. 4 shows *F*-measures for the SVM detectors trained on Weekly Reader data only and Weekly Reader plus newswire data, and SVM regression classifiers trained on both datasets. When trained on the Weekly Reader data, the SVM regression classifier performs comparably to the SVM detectors for grades 2 and 3, slightly worse on grade 4, and slightly better for grade 5. However, the regression classifier trained on Weekly Reader plus newswire data has worse performance than the detectors at all grade levels except grade 3. There are a few possible reasons for this. First, it may be problematic to use a single grade level for the negative training data, in which case additional labeled training data might be useful. Second, the error model assumes a linear distance function from 2 to 9, and it may be that a nonlinear mapping makes more sense, e.g., the differences between 5 and 9 may not be much greater than the differences between 2 and 4.

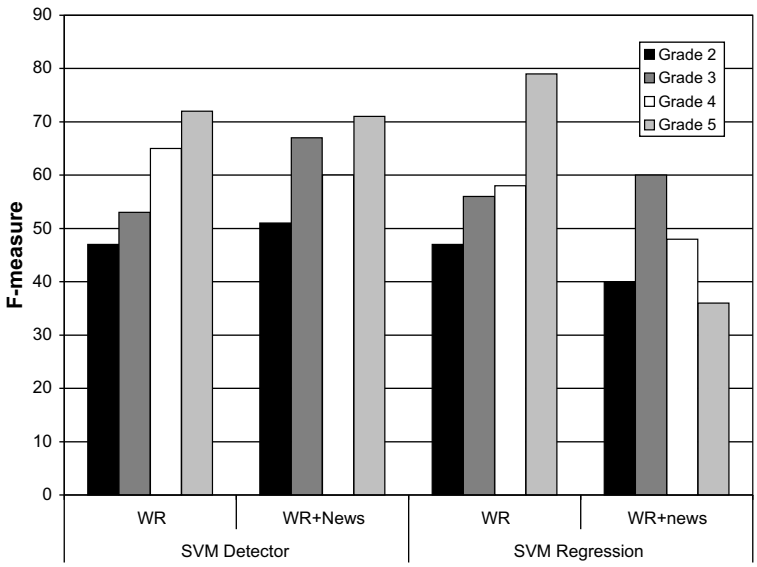


Fig. 4. Comparison of SVM detectors and SVM regression classifier.

Since tasks such as annotating articles on the web will not have fixed upper limits on the articles scored, the advantage in generalization performance on other data is of substantial real-world importance. Hence, in all subsequent experiments, we train SVMs on Weekly Reader plus newswire data.

6.4. SVM feature analysis

Our prior work showed that the SVM with language model features outperforms a language model alone, but within the SVM framework it may be that only very simple lexical features are needed to achieve good performance. Hence, we investigated the degree to which the syntactic features explored here provide a benefit over traditional features. We divided the features into the following groups:

- *lexical-only*: OOV and average number of syllables per word,
- *syntactic*: parse and n-gram scores (unigram, bigram, trigram),
- *non-syntactic*: all lexical-only features plus average sentence length in words and Flesch-Kincaid score.

The syntactic features represent the new information used in our approach, and the non-syntactic features correspond to the traditional approach. The lexical-only features omit sentence length (and thus Flesch-Kincaid which incorporates sentence length), and they are included to assess the relative importance of vocabulary vs. structural cues.

We trained new versions of the SVM grade-level detectors with each of the above categories of features. Fig. 5 shows *F*-measures for these classifiers compared to the classifiers using all features, trained on the Weekly Reader training set augmented with newswire data. The SVMs trained with lexical features perform comparably to the SVMs trained with all features for grades 2 and 3, while for the higher grades, the classifiers that use all types of features give better results. The SVMs trained with syntactic features alone do not perform as well as the other classifiers, but these features still appear to contribute to the overall performance of the SVMs trained with all features.

To study the relative importance of the four parser features, we trained a decision tree classifier (C4.5, Quinlan, 1993) with the Weekly Reader training set using only the parser features to classify articles in grades 2 through 5. Our goal was not to use this classifier explicitly to do grade-level classification, but to see how it made use of the four features. All four parse features were used by the decision tree. The features for the

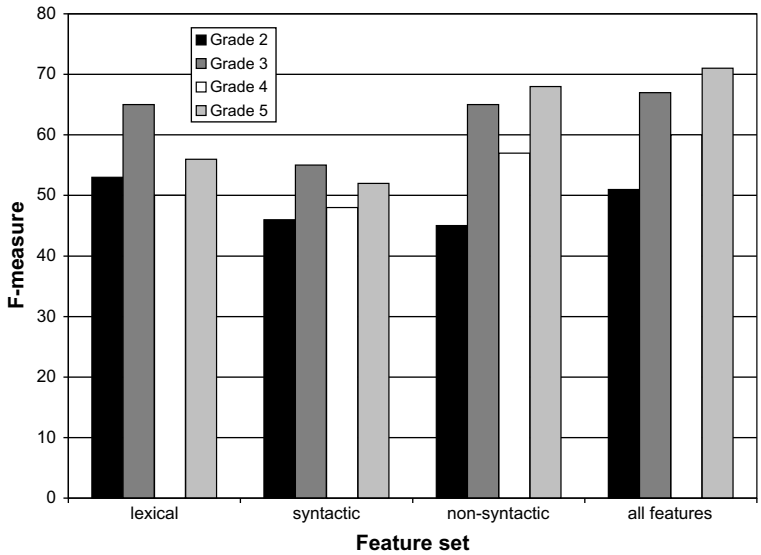


Fig. 5. Comparison of *F*-measures for SVM detectors trained with lexical, syntactic and non-syntactic features on the Weekly Reader data, covering grades 2 through 5, plus negative examples from TIPSTER newswire data.

average number of noun phrases and verb phrases were used at higher nodes in the tree, while the features for the average number of SBARs and the average parse tree height were used for more fine-grained decisions at lower nodes of the tree.

6.5. Comparison with other methods

We compared the regression and detection SVMs with two traditional reading level measures, Flesch-Kincaid and Lexile, chosen because of their popularity and the availability of tools for computing the scores. (The Lexile tool is available online but had to be run manually for each passage.) The Flesch-Kincaid score for a document is intended to directly correspond with its grade level, which we rounded to get an integer level in our experiments. The Lexile scores do not correlate directly to numeric grade levels, but a mapping to the corresponding grade levels is available on the Lexile web site (The Lexile framework for reading, 2005).

Since these numbers correspond to classifiers (vs. detectors) performance can be evaluated in terms of accuracy or *F*-measure. The accuracy of the Flesch-Kincaid index is only 5%, while Lexile's accuracy is 36% and the SVM detectors achieve 43%. Fig. 6 shows *F*-measures for the Flesch-Kincaid and Lexile measures compared to the two SVMs trained on Weekly Reader plus newswire data. 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. Both the SVM regression classifier and detector outperform the Lexile metric.

A problem with *F*-measure is that all errors are counted as equal. Since an error of one grade level may be acceptable, we also compare the different reading level classifiers in terms of what percentage of articles are classified at a level that is more than one different from the Weekly Reader classification. The results in Table 6 show the same pattern of significantly better performance achieved with the SVM.

Lexile is a more general measure while our regression classifier is trained on this particular domain, so the better performance of our model is not entirely surprising. Importantly, however, our classifier is easily tuned to any corpus of interest. In addition, we find in the next section that the SVM detector better matches human labels that are not tuned to the Weekly Reader definition.

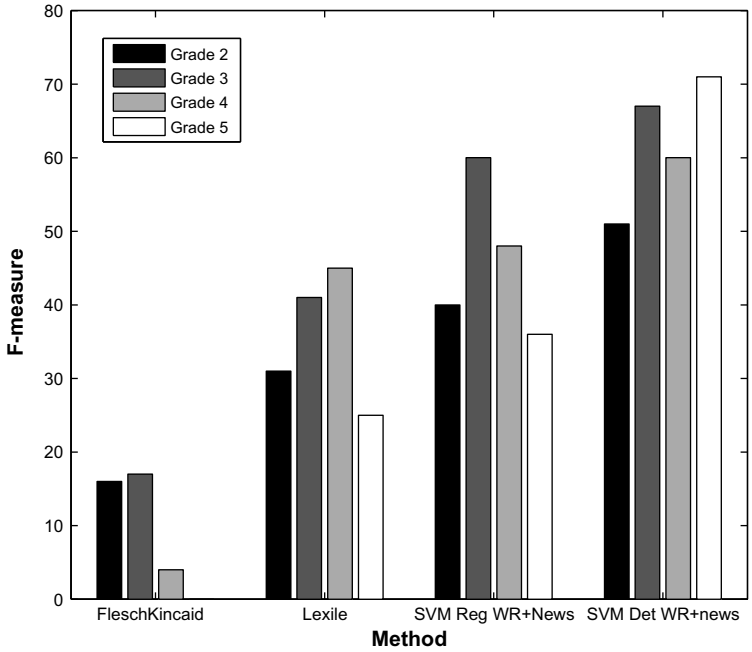


Fig. 6. Comparison of Flesch-Kincaid, Lexile, SVM regression classifier and SVM detectors trained on Weekly Reader plus news data.

Table 6
Percentage of articles which are misclassified by more than one grade level by traditional and SVM classifiers

Grade	% errors > 1level		
	Flesch-Kincaid (%)	Lexile (%)	SVM (%)
2	78	33	0
3	67	27	3
4	74	26	13
5	59	24	9

7. Assessment with multiple annotations

One of the challenges in the area of reading level assessment is knowing the right answer. In the experiments described above, 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, how well human annotators agreed with each other and with the labels given in the corpus, and how well our detectors perform when compared to human evaluations of this 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.

To investigate this issue, we conducted a study of the performance of human annotators on the Weekly Reader data. We hired three experts to annotate our test corpus, including an elementary school bilingual education teacher (annotator A) and two graduate students in fields relevant to reading and pedagogy (annotators B and C). We provided the annotators with a few 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 (i.e., a single article can have hits from more than one detector).

We measured human variability in three ways: Cohen's kappa statistic for measuring inter-rater agreement (Carletta, 1996), the percentage of documents on which annotators disagreed by more than one level, and the *F*-measure associated with comparing human labels to the Weekly Reader labels (in the same way as for our SVM detectors). We also illustrate the per-document agreement in radar plots.

The kappa value is calculated by comparing binary annotations from a pair of labelers, where the annotations indicate whether or not an article matches a particular grade level. The kappa values observed for the three different pairs of labelers ranged from 0.40 to 0.54, which indicates moderate but not good agreement (which would correspond to a kappa of 0.6–0.7). The kappa for the Weekly Reader labels vs. the human labelers is much lower. The kappa for the Weekly Reader labels vs. the SVM is in the same range: 0.52.

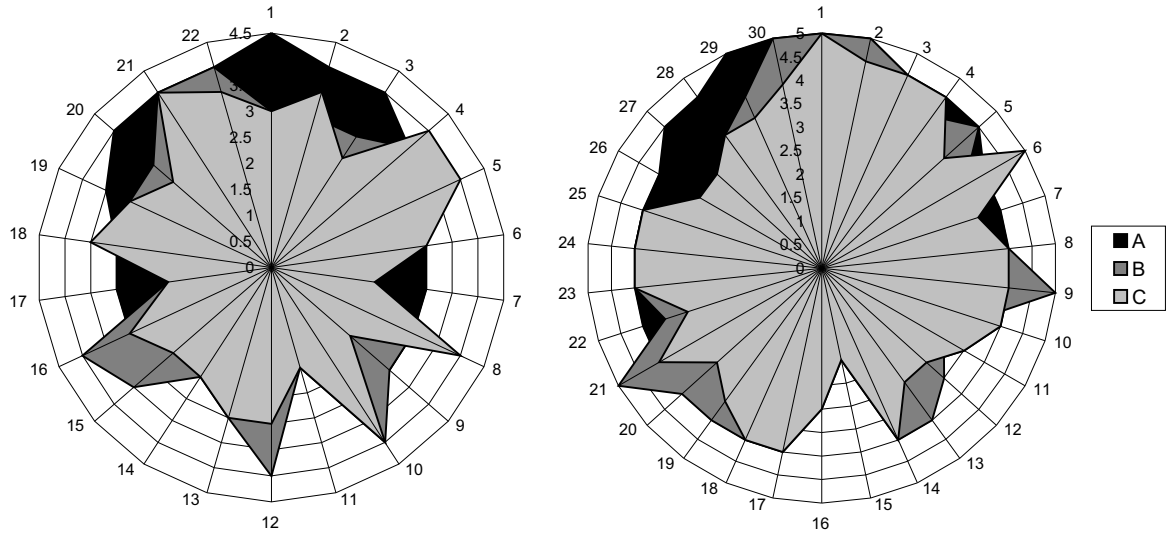
The kappa measure does not capture the fact that a disagreement in grade levels of 2 or 3 is worse than a disagreement of 1. Hence, we also looked at the percentage of articles where the humans disagree by more than one grade-level, and found that this happens for 26% of the articles.

Interannotator agreement for documents of grade levels 3 and 4[4] is illustrated with the radar plots in Fig. 7, following the presentation introduced in Kolluru et al. (2003). Each radial line corresponds to a document, and the annotators are represented with different shades of gray. Documents where annotators thought that two grade levels were reasonable matches are indicated with the average of the two, e.g., 4.5 for grades 4 and 5. Note that the disagreement among humans is not often greater than 1.

Finally, we look at *F*-measure for the annotators' labels evaluated on the Weekly Reader test set labels. The human annotators are treated as "detectors" and evaluated in the same way as the SVMs. Fig. 8 shows the *F*-measure results for the labels provided by annotators A, B and C compared with results for the SVM detectors.[5] We observe that the three human annotators have roughly similar performance, but all of the

[4] Grades 3 and 4 are most informative, since they do not benefit from being at the ends of the scale (for this data).

[5] The annotation experiments used a subset of about 80% of the original test set. The SVM and Lexile results in this figure are for this subset only and do not exactly correspond to the results in Section 6.5.



(a) Grade 3 (b) Grade 4

Fig. 7. Radar plots showing interannotator agreement for documents at grade levels 3 and 4.

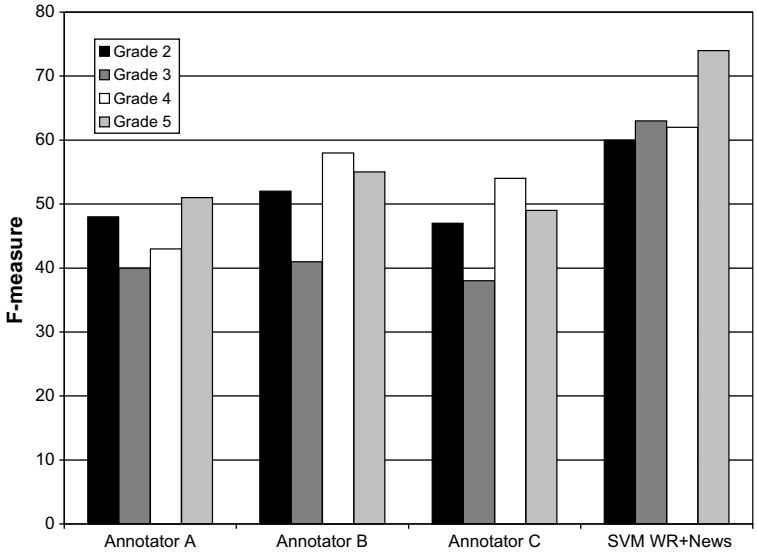


Fig. 8. *F*-measures for annotators A, B, and C compared with SVM detectors trained on Weekly Reader plus newswire data.

numbers for the human annotators are less than 60%, some much lower. We also find that the SVM detectors (being trained specifically for the Weekly Reader task) have higher agreement with the Weekly Reader labels than the human agreement with these labels. This is not to say that machines outperform humans in general for reading level detection, but rather that they are able to learn something about the Weekly Reader conventions.[6]

What all of these measures show is, first, that there is a fair amount of disagreement among the human annotators, even for people with appropriate education and preparation for the task. It is likely that conven-

[6] The criteria used by Weekly Reader are not published, and we were not able to obtain details from them.

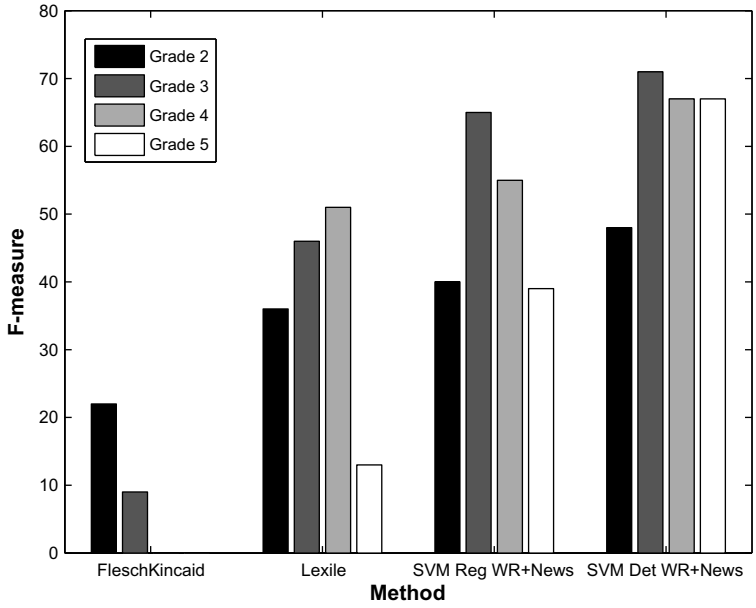


Fig. 9. Comparison of Flesch-Kincaid, Lexile, and SVM regression classifier and SVM detectors trained on Weekly Reader plus news data and evaluated on a subset of 73 articles on which annotators do not disagree by more than one grade level.

tions people use are influenced by their audience (readership for a magazine, students that a teacher works with). The SVM agreement with the Weekly Reader labels is better than the human agreement with these labels, but the humans agree among themselves more reliably than with Weekly Reader, so we conjecture that our annotators are using somewhat different criteria. This effect is particularly notable for grade 3, but it may be true for lower grades in general and not measurable with our data since the human annotators knew that 2

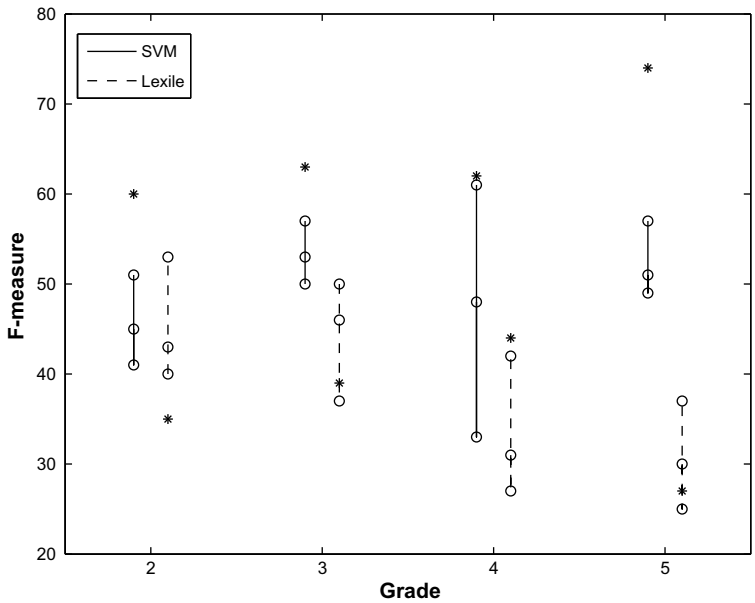


Fig. 10. *F*-measures for Lexile and SVM detectors trained on Weekly Reader and newswire data compared to labels provided by human annotators. Asterisks indicate the *F*-measure for each classifier with respect to the Weekly Reader test set labels.

was the lowest level. The SVM better matches the Weekly Reader labels, because it is trained on other Weekly Reader articles. These findings support the use of machine learning for this task, since it is possible to tailor reading level detectors to the needs of a particular student or group of students.

If we assume that the articles where humans disagreed by more than one level are either especially difficult or possibly idiosyncratic, then it is interesting to compare performance of different reading level assessment tools only on the subset of articles where humans agreed. The results are shown in Fig. 9. On this "cleaner" test set, the results are a bit higher for all classifiers, but the trends remain the same.

Since the human labels appeared to be somewhat different from the Weekly Reader labels, it is interesting to compare the performance of the SVM detectors and the Lexile measure using the labels provided by the human annotators as a gold standard. We calculate *F*-measure for the SVMs and Lexile compared to each human annotator's labels individually and present the results in Fig. 10. In the figure, lines connect the classifier scores for each grade relative to all three annotators. Asterisks indicate the *F*-measure for each classifier evaluated on the Weekly Reader test set labels. The Lexile results compared to human annotators tend to fare better than when compared to the Weekly Reader test set labels (except for grade 4), and the SVM tends to have somewhat worse performance, consistent with the fact that Lexile is a general classifier whereas the SVM is tuned to the Weekly Reader corpus. However, the SVM results are still almost always higher than the Lexile results. Thus, the Weekly Reader data and the data-driven learning paradigm is reasonable even for a more general user community.

8. Conclusions and future work

In summary, 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 unlabeled negative training data can be used to augment a corpus with only positive labels, but more effectively for a detection paradigm than regression. We also confirm other findings that syntactic features have only a small effect on the overall performance of the detectors. Both regression and detection SVMs compare favorably to other existing methods using several different methods to measure performance. The SVM performance is better than that of human annotators when both are compared to the Weekly Reader labels, but the humans appear to be using different annotation criteria since their inter-annotator agreement is higher than their agreement with the Weekly Reader labels.

There are many possibilities for improving performance through more extensive exploration of model structures and feature extraction techniques, including using different size OOV lists, new parse features, different feature selection thresholds for the class language models, etc. However, our initial experiments in extending the feature vector showed limited success probably due to the small amount of training data available, despite the reputation of SVMs being able to handle large dimensional feature vectors.

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 specific needs of particular students. The variability in human annotation also supports a need for automatic learning, which can tune to the particular conventions that specific teachers want to use. Since annotated data is so difficult to find, development of adaptation techniques will be important. Some possible directions include SVM adaptation using active learning, e.g., Tong and Koller (2001) or relevance feedback, e.g., Drucker et al. (2001).

Acknowledgements

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

Barzilay, R., Elhadad, N., 2003. Sentence alignment for monolingual comparable corpora. In: Proc. of EMNLP, pp. 25–32.

Bylsma, P., Ireland, L., Malagon, H., 2003. Educating English Language Learners in Washington State. Office of the Superintendent of Public Instruction, Olympia, WA, 2003.

Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics 22 (2), 249–256.

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

Charniak, E., 2000. A maximum-entropy-inspired parser. In: Proc. of NAACL, pp. 132–139.

Chen, S., Goodman, J., 1999. An empirical study of smoothing techniques for language modeling. Computer Speech and Language 13 (4), 359–393.

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

Damashek, M., 1995. Gauging similarity with n-grams: language-independent categorization of text. Science 267 (5199), 843–848.

Drucker, H., Shahraray, B., Gibbon, D.C., 2001. Relevance feedback using support vector machines. In: Proc. of ICML, pp. 122–129.

Fisher, W., 1996. A C implementation of Daniel Kahn's theory of English syllable structure. <ftp://jaguar.ncsl.nist.gov/pub/tsylb2-1.1.tar.Z>.

Fountas, I.C., Pinnell, G.S., 1999. Matching Books to Readers: Using Leveled Books in Guided Reading K-3. Heinemann, Portsmouth, NH.

Gunning, R., 1952. The Technique of Clear Writing. McGraw-Hill, New York.

Harman, D., Liberman, M., 1993. TIPSTER Complete. Linguistic Data Consortium, catalog number LDC93T3A and ISBN: 1-58563-020-9, 1993.

Hellman, M., Collins-Thompson, K., Callan, J., Eskenazi, M., 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In: Proc. of the NAACL/HLT Conference, 2007, pp. 460–467.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A practical guide to support vector classification. <http://www.csie.ntu.edu.tw/cjlin/papers/guide/guide.pdf> (Accessed 11/2004).

Huffman, S., 1995. Acquaintance: language-independent document categorization by n-grams. In: Proc. of TREC-4, 4th Text Retrieval Conference, pp. 359–371.

Joachims, T., 1998. Text categorization with support vector machines: learning with many relevant features. In: Proc. of the European Conference on Machine Learning, pp. 137–142.

Joachims, T., 1999. Making large-scale support vector machine learning practical. In: Schölkopf, B., Burges, C., Smola, A. (Eds.), Advances in Kernel Methods: Support Vector Machines. MIT Press, Cambridge, MA.

Kessler, B., Nunberg, G., Schütze, H., 1997. Automatic detection of text genre. In: Proc. of ACL/EACL, pp. 32–38.

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

Kolluru, B., Christensen, H., Gotoh, Y., Renals, S., 2003. Exploring the style-technique interaction in extractive summarization of broadcast news. In: Proc. of the Automatic Speech Recognition and Understanding Workshop, pp. 495–500.

Lee, Y.-B., Myaeng, S.H., 2002. Text genre classification with genre-revealing and subject-revealing features. In: Proc. of SIGIR, pp. 145–150.

Peng, F., Schuurmans, D., Wang, S., 2003. Language and task independent text categorization with simple language models. In: Proc. HLT-NAACL, pp. 110–117.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

Ratnaparkhi, A., 1996. A maximum entropy part-of-speech tagger. In: Proc. of EMNLP, pp. 133–141.

Schwarm, S.E., Ostendorf, M., 2005. Reading level assessment using support vector machines and statistical language models. In: Proc. of ACL, pp. 523–530.

Si, L., Callan, J., 2001. A statistical model for scientific readability. In: Proc. of CIKM, pp. 574–576.

Stamatatos, E., Fakotakis, N., Kokkinakis, G., 2000. Text genre detection using common word frequencies. In: Proc. of COLING, pp. 808–814.

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

Stolcke, A., 2002. SRILM – an extensible language modeling toolkit. In: Proc. of ICSLP, vol. 2, pp. 901–904.

The Lexile framework for reading, 2005. <http://www.lexile.com> (Accessed April 15, 2005).

The Washington Post, 2005. <http://www.washingtonpost.com> (Accessed April 20, 2005).

Tong, S., Koller, D., 2001. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research 2 (Nov), 45–66.

US Department of Education, 2003. National Center for Educational Statistics. NCES fast facts: Bilingual education/Limited English Proficient students. <http://nces.ed.gov/fastfacts/display.asp?id=96> (Accessed June 18, 2004).

US Department of Education, 2005. National Center for Educational Statistics. The condition of education 2005. <http://nces.ed.gov/pubs2005/2005094.pdf> (Accessed November 17, 2005).

Weekly Reader, 2004. http://www.weeklyreader.com (Accessed July, 2004).

Western/Pacific Literacy Network/Literacyworks, 2004. CNN SF learning resources. <http://literacynet.org/cnnsf/> (Accessed June 15, 2004).

Yamron, J.P., Carp, I., Gillick, L., Lowe, S., van Mulbregt, P., 1999. Topic Tracking in a News Stream. In: Proc. of the DARPA Broadcast News Workshop, pp. 133–136.
Yang, Y., Pedersen, J., 1997. A comparative study on feature selection in text categorization. In: Proc. ICML, pp. 412–420.