Assessing the Reading Level of Web Pages

Sarah E. Petersen, Mari Ostendorf

Dept. of Computer Science, Dept. of Electrical Engineering
University of Washington
Seattle, WA 98195, USA
sarahs@cs.washington.edu, mo@ee.washington.edu

Abstract

Reading is an important part of educational development. However, finding appropriate reading material for all students can be difficult and time consuming for teachers. Our goal is to automate the task of assessing the reading level of text to enable teachers to more effectively take advantage of the large amounts of text available today on the World Wide Web. Reading level assessment tools already exist for clean corpora such as books and magazine articles. This paper presents extensions of a particular set of tools to handle web pages returned by a standard search engine, including a step that pre-filters web pages to eliminate "junk" pages with little or no text. Results of applying the reading level detectors to web pages are manually evaluated by elementary school teachers, the intended audience for these tools. The tools work well for grades 4 and 5, with room for improvement in grades 2 and 3.

Index Terms: reading level assessment, SVMs, web pages.

1. Introduction

Reading is a key component of language and educational development, and it is important to provide appropriate reading material for each student. This goal is particularly challenging in bilingual education and English as a Second Language (ESL) settings, in which a student's reading ability in English usually does not match his or her intellectual ability in general, and there may be a wide variety of reading abilities among students in the same classroom. Students whose first language is not English form a large and growing group in the United States [12].

Teachers must find appropriate reading material at a variety of levels for their students to read independently or with assistance. For Limited English Proficient students, there is an added challenge: teachers seek "high interest level" texts at low reading levels to meet the needs of students who read below their age-appropriate grade level. Graded textbooks and other materials are usually available, but these do not always meet the high interest/low reading level criterion. Additionally, students often need to do supplemental reading outside of graded textbooks for class projects.

We would like to leverage the World Wide Web, with vast amounts of text on many topics, as a source of supplemental texts for teachers and students; the difficulty is efficiently sifting through these texts. Searching the web by topic is easily accomplished by search engines such as Google, but for educational needs it is also important to filter the resulting web pages by grade level. Teachers do not always have time to sift through many pages of search engine results to find web pages at the right reading level for their students. The goal of this work is to automate this task to enable teachers and students to make better use of texts available on the Web. This paper describes our approach to augmenting tools that were previously developed on a clean corpus to handle texts from the web. The motivating scenario for this work is a teacher or student looking for web pages on a particular topic, e.g., for a research project, at a specified grade level. We use topics such as animal names, countries, and natural phenomena (e.g. tornadoes). We developed this scenario and these topic lists in consultation with teachers at a local elementary school with bilingual education and ESL programs. Our approach combines our previous work on reading level detection with ideas drawn from web classification results by Sethy et al. [9].

The rest of the paper is organized as follows. Section 2 provides background on reading level assessment and describes the tool used in this work. In section 3, we describe a pilot study on this task and discuss lessons learned. Section 4 presents our approach to pre-filtering the web data, and section 5 presents results for the reading level detectors applied to filtered web data. Section 6 summarizes our findings and describes future plans for customization of the reading level detectors for individual users.

2. Reading Level Detection

In traditional approaches to reading level assessment, the focus is often on easy-to-calculate approximations of semantics and syntax. For example, the Lexile framework [10] measures semantics by word frequency counts and syntax by sentence length. In more recent work, Collins-Thompson and Callan improve upon the traditional classifiers with a "smoothed unigram" classifier which better captures the variance in word usage across grade levels [2]. This reading level classifier is used in REAP, a system designed to select reading material from the web for students in second-language classes based on unigram models of grade level, curriculum, and the reading level of individual students [1].

The REAP approach reflects an emphasis on vocabulary acquisition, which is facilitated by the use of unigram models. In some contexts, it is also useful to look at short phrases and/or the complexity of syntactic structure in a text. In particular, this may be important for the ESL task, where the vocabulary (i.e., topic) and grade level are not necessarily well-matched and teachers want materials with more difficult, topic-specific words but simple structure. In [8, 6], we showed that further gains can be achieved by using higher-order n-grams and incorporating syntactic features from an automatic parser. We pose the reading level assessment problem as a detection task, with one detector per grade level which decides whether or not a particular article belongs to that grade level category. The detectors are support vector machine

(SVM) classifiers, with features that include traditional grade level features (e.g., average number of words per sentence), n-gram language model scores, and parser-based features. Experiments are conducted with texts from Weekly Reader, an educational newspaper with versions targeted at different grade levels [13]. Weekly Reader articles cover topics such as current events, science, and history. Detectors trained for grade levels 2-5 outperform the traditional Flesch-Kincaid and Lexile reading level measures, with F-measures in the range of .5-.7 (depending on grade level) for data within the 2-5 range. Text at adult reading levels is rejected in 90% of the 30 cases that we tested.

In this paper, we address the problem of moving from a static collection of good quality text to the dynamic (but mixed quality) text resources found on the web, with the goal of online access for teachers and students. We assume that the search engine (Google) is doing a good job on finding topic-relevant pages, constraining the focus of this effort to filtering the results to find text at the appropriate reading level. In the pilot study described next, we find that the web pages returned include a large number of pages that the detectors trained on clean text simply are not designed to handle. Hence, the tools need to be augmented to first distinguish web pages with narrative text from those that mainly have links, advertisements, or other unwanted content. Thus, this study looks at performance of the existing reading level detectors [6] on raw web text, as well as text filtered to remove "junk" automatically using additional tools developed here.

3. Pilot study

We conducted a pilot study in which we applied our grade level detectors to web pages returned by Google for several topics and presented the top 15 positive results in each category to two elementary school bilingual education teachers.[1] The topics were elephants, giraffes, Japan and Mexico. The queries we submitted to Google consisted of the query term and the word "kids" with the intention that this would help find pages that were geared towards children. In some cases this was successful, but in other cases it returned an excessive number of sales pages with no real content, e.g., sites selling kids t-shirts with pictures of elephants on them. In subsequent experiments, we used the topic term only.

After retrieving topical web pages from Google, we used simple heuristic filters to clean up the text. We removed HTML tags and filtered the articles to keep only contiguous blocks of text consisting of sentences with an out of vocabulary word (OOV) rate of less than 50% relative to a general-purpose word list of 36k English words. Then we ignored articles without at least 50 words remaining and applied the reading level detectors described in the previous section to the remaining texts. Finally, we sorted the results by the score given by each grade level detector and presented the top 15 pages for each topic and grade to the annotators. In some cases, the total number of articles was less than 15. Despite downloading a large number of articles for each topic, we discovered that some topic/grade combinations did not result in many hits for the grade level classifiers.

Annotators were asked to view each page and choose from the following labels:

- Good = Acceptable for this grade level.
- Too low = Too easy for this grade level.
- Too high = Too high for this grade level.
- N/A = Off topic or not appropriate for students.

The annotators were not trained on sample texts for each grade level, with the premise that eventual users of the system (teachers) will not want to have special purpose training (other than their education backgrounds). From our prior work [6], we know that there are considerable individual differences between teachers, which should be respected, so the goal here is to improve the percentage of useful texts returned between different trials for the same person.

Table 1: *Pilot study teacher annotation results. Annotator A viewed 151 articles, and Annotator B viewed 135 articles.*

	Percentage of articles	
Label	Annotator A	Annotator B
Good	24%	35%
Low	1%	0%
High	5%	7%
N/A	69%	58%

The annotators viewed the topics in a different order from each other, and neither finished all the articles in the time allotted for the pilot study. Table 1 shows the percentage of articles they annotated with each label. Clearly, these results indicate much room for improvement. Most articles are off topic or otherwise inappropriate. While the categories given to the annotators did not allow us to distinguish between these reasons, anecdotal examples suggested that we needed to improve our filtering techniques to remove "junk" pages and increase the quality of the texts submitted to the reading level detectors. We also observed that sorting the web pages by the score of the reading level detector had the unfortunate side effect of returning articles that happened to score well for reading level but were far down the original list of search engine hits and likely to be off topic. In order to increase the topicality of the results, in subsequent experiments the order of pages returned by the search engine is respected. We step through the pages in their original order according to the search engine, returning the first N pages that are classified positively by the reading level detector.

4. Filtering web text

4.1. Heuristics

Based on observations in the pilot study, we made some changes to improve the heuristics used to filter web pages prior to applying the reading level detectors. We continue to filter sentences with an OOV rate greater than 50% but no longer stipulate that all remaining sentences in an article must be from a contiguous chunk. We also filter lines with fewer than 5 words; this removes many leftover formatting and navigation elements of the original web page such as menu bar items, titles, and other web-page-specific artifacts that are not part of the main text of the page.

4.2. Naive Bayes classifier

Unfortunately, the heuristics described above eliminate only a small portion of the non-applicable articles. Many web pages returned by the original Google query are "junk", containing little text to be classified. Inspired by Sethy et al.'s work on web page classification [9], we designed a naive Bayes classifier to distinguish "content" pages from "junk." We select features for this

[1] These teachers served as annotators for all the reading level experiments described in this paper.

classifier according to their information gain (IG) [14]. Information gain measures the difference in entropy when w is and is not included as a feature:

$$\begin{aligned} IG(w) = & - \sum_{c \in C} P(c) \log P(c) \\ & + P(w) \sum_{c \in C} P(c|w) \log P(c|w) \\ & + P(\bar{w}) \sum_{c \in C} P(c|\bar{w}) \log P(c|\bar{w}), \end{aligned} \quad (1)$$

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 part-of-speech tag, as labeled by a maximum entropy tagger [7]. After selecting features, a naive Bayes classifier was trained with the *rainbow* program from the Bow toolkit [5].

For our original version of the content vs. junk classifier, we used the annotated web pages from the pilot study as training data. There were 86 negative examples (i.e., "junk" pages) and 36 positive examples (i.e., "content" pages) which we augmented with 25 additional hand-selected pages for a total of 61 positive examples. The feature selection process described above resulted in 404 word features for the original naive Bayes classifier. The number of features increased somewhat in further iterations, described next, but remained in the range of 400 to 500 words.

In order to validate and improve this classifier, we conducted two iterations of annotation and retraining. This technique of selecting additional examples for annotation to improve the classifier is similar to both active learning, as described in the context of SVMs in [11], and relevance feedback, as presented for SVMs in [3]. In the active learning scenario, the examples for which the classifier is most uncertain are presented to the annotator for feedback. In the relevance feedback approach, the user provides feedback on the highest ranked examples. In our work, articles are ranked topically by the search engine, and we choose the first N which are positively classified by the content vs. junk classifier for annotation. These are not necessarily either the most certain or most uncertain examples according to the content vs. junk classifier alone, but they do correspond to the highest ranked examples of the system overall, combining topic and content/junk classification, which is in fact what we want to improve.

Specifically, at each iteration we presented an annotator with a set of articles from each of eight topic categories; the categories were different in each iteration and different from the topics used in the pilot study. Once again, the topics were names of animals, countries, and weather phenomena. The articles presented to the annotator were the first 30 positively classified articles for each topic according to the content vs. junk classifier. We sought to optimize our use of the annotator's time by only showing them the articles that the classifier believed were "content" and having them confirm or correct this classification. After the first iteration, we added the newly annotated data to the training data and re-ran the feature selection and classifier training steps. We did the same after the second iteration, resulting in the third-iteration content vs. junk classifier used to filter web pages for the reading level detection experiments described in the next section. Table 2 shows the number of examples of both content and junk pages used to train each iteration of the content vs. junk classifier. From this data, one can see that an increasing percentage of good content pages are being returned with each iteration, though the numbers are not strictly comparable because the samples are different.[2]

Table 2: *Number of training samples for each iteration of the content vs. junk classifier.*

Iteration	Num Content	Num Junk	% Content
1	61	86	.41
2	217	159	.58
3	361	241	.60

Table 3: *Percentage correct classification for content vs. junk classifier at each iteration and for each category.*

Iteration	Content Correct	Junk Correct	Total Correct
1	80%	70%	75%
2	94%	55%	75%
3	93%	57%	75%

To compare performance across all three iterations of the content vs. junk classifier, we applied these classifiers to a third set of web pages consisting of approximately 60 pages per topic for each of six new topics. The human annotator labeled these pages as either content or junk, resulting in a test set of 348 pages, of which 177 were content and 171 were junk. We applied the three versions of the content vs. junk classifier to this test set. Table 3 shows the percentage of articles which were correctly classified in each category. Note that the total percentage of correctly classified pages is the same across all three iterations, but substantially more content articles are classified correctly by the second and third iteration classifiers. This comes at the expense of more junk articles that are incorrectly classified (i.e., false positives), though some of these may be elimated in the reading level detection stage.

5. Reading level experiments

We conducted a set of reading level detection experiments on six novel topics (rainforests, hurricanes, tsunami, frogs, leopards and owls) with the same teachers serving as annotators as in the pilot study. We downloaded approximately 1,000 web pages per topic, using the topic word as the Google query. These pages were filtered using the heuristics and the naive Bayes classifier described in the previous section, resulting in between 325 and 450 pages on each topic. The reading level detectors described in section 2 were applied to these articles. The first 10 hits per topic that were classified positively by each grade level detector were presented to the annotators for labeling. Many topics did not have a full 10 hits for grade 2; this is probably due to most web pages being at higher reading levels.

The annotators were asked to choose from the following set of labels for each article. These are similar to the labels used in the pilot study, with slight changes based on observations from that study. "Just links" refers to web pages which provide links to other pages but do not have much content in terms of paragraphs of text to read. These pages are not necessarily off topic, but they are also not the sort of text on which we hope to detect grade level. Anno-

[2] In fact, the gain is bigger than indicated in the table, since the actual number of positive content hits initially was 36, or 30% content.

tators were also able to provide free response comments about any page about which they wished to make an additional note.

- Good = Acceptable for this grade level.
- Too low = Too easy for this grade level.
- Too high = Too high for this grade level.
- Just links = A page of links with no significant text. Probably still on topic.
- Off topic.

Table 4: *Summary of annotations for web pages. Percentages do not sum exactly to 100% because in some cases, the annotators marked more than one label, e.g., good reading level but off topic.*

	Percentage of articles	
Label	Annotator A	Annotator B
Good	59%	41%
Low	0%	0%
High	8%	18%
Just links	14%	16%
Off topic	20%	21%

Table 4 shows the overall percentage of labels selected by the annotators for all the web pages they viewed. Neither annotator thought that any of the articles detected for any grade level were too low for that level. This pattern of detecting articles that are too high but not too low is different from that observed on the "clean" Weekly Reader data, probably because the material on the web is not balanced for reading level and has fewer low-grade-level articles. Annotator B found more articles that were too high than annotator A did; as seen earlier, there are individual differences in judgments about reading level. However, both annotators labeled a significantly larger percentage of the data as "good" and a much smaller percentage in the not applicable categories. In addition, the percentages of pages labeled "just links" and "off topic" are similar for the two annotators. Interestingly, the annotators indicated in free response comments that the off topic articles were off topic but not inappropriate for kids, e.g. a web page for a research study with the acronym OWLS for its name, and a web page about the town of Owlshead both appeared for the query "owls."

Table 5: *Percentage of articles of each grade level labeled "good" by each annotator.*

Grade	Annotator A	Annotator B
2	40%	30%
3	37%	23%
4	66%	47%
5	81%	56%

Table 5 shows the percentage of web pages marked as "good" for each grade level by the two annotators. In general, the reading level detectors are more accurate for the higher grade levels.

6. Conclusions and future work

Our SVM-based reading level detectors trained on Weekly Reader text can be applied successfully to text from web pages. However, filtering the web pages to eliminate "junk" pages before applying the reading level detectors is essential. In addition, the high percentage of off-topic pages suggest that more strict topic filtering may be useful, perhaps by using fewer of the returned pages.

Reading level assessment is a subjective problem. Results vary between annotators, as different users have different perceptions of the appropriateness of articles for a particular grade level. We will to address this issue in future research by using annotations from each teacher to adapt the reading level detectors to better meet the needs of each individual user.

7. Acknowledgments

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

8. References

[1] Brown, J., and Eskenazi, M., "Retrieval of Authentic Documents for Reader-Specific Lexical Practice", Proc. INSTIL, 2004.

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

[3] Drucker, H., Shahrary, B. and Gibbon, D., "Relevance feedback using support vector machines", Proc. ICML, 122-129, 2001.

[4] Joachims, T., "Making large-scale support vector machine learning practical", Advances in Kernel Methods: Support Vector Machines, B. Schölkopf, C. Burges, A. Smola, eds. MIT Press, Cambridge, MA, 1999.

[5] McCallum, A. K., "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering." http://www.cs.cmu.edu/ mccallum/bow. 1996.

[6] Petersen, S. E. and Ostendorf, M., "A machine learning approach to reading level assessment," University of Washington CSE Technical Report 2006-06-06.

[7] Ratnaparkhi, A., "A maximum entropy part-of-speech tagger", Proc. of EMNLP, 133–141, 1996.

[8] Schwarm, S. E. and Ostendorf, M., "Reading level assessment using support vector machines and statistical language models", Proc. ACL, 523-530, 2005.

[9] Sethy, A., Georgiou, P. G., and Narayanan, S., "Building topic specific language models from webdata using competitive models", Proc. INTERSPEECH, 1293-1296, 2005.

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

[11] Tong, S. and Koller, D., "Support vector machine active learning with applications to text classification", Journal of Machine Learning Research, 2(Nov):45-66, 2001.

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

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

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