Read-X: Automatic Evaluation of Reading Difficulty of Web Text

Eleni Miltsakaki
Graduate School of Education
University of Pennsylvania, USA
elenimi@gse.upenn.edu

Audrey Troutt
School of Engineering and Applied Science
University of Pennsylvania, USA
atroutt@seas.upenn.edu

Abstract: We are developing a web-search application to locate and evaluate potential reading material on the internet. Our application, Read-X, performs a keyword search of the internet, analyzes the readability of text from each resulting website and classifies the text according to theme. This tool will be useful to adolescent and adult low-level reading students who face, among other challenges, a troubling lack of reading material for their age, interests and reading level.

Introduction

According to the National Center for Education Statistics, 29% of high school seniors in public schools across America were below basic achievement in reading in 2005 (U.S. Department of Education 2005). Once these students enter high school, their reading problems, which began much earlier in their education, are compounded by many factors including a lack of suitable reading material for their age, interests and reading level. Most material written at a lower reading level is designed for much younger students; high-school students find it boring or embarrassing. On the other hand material designed for older students, while probably more interesting, is incomprehensible to such a student and leads to frustration and self-doubt. The internet is a vast resource for potential reading material and is often utilized by educators in the classroom, but it is not currently possible to filter the results of a search engine query by levels of readability. Instead, the software that some schools have adopted restricts students to lists and directories of hand-selected educational sites. This severely limits the content available to students and requires near-constant maintenance to keep current with new information available on the web.

We are developing a web-search application to locate and evaluate text on the internet. Our application, Read-X, analyzes the readability of each text and can filter the results according to readability for the user. This tool will be useful to adolescent and adult low-level reading students who face, among other challenges, a troubling lack of reading material for their age, interests and reading level

The remaining of the paper is organized as follows: first we will describe our motivation for creating Read-X, which is based on studies that show that older struggling readers can make improvements in literacy and that those improvements can have a profound impact on their lives. Next we will describe existing technologies for literacy improvement and research related to our current project. Finally, we will give a detailed description of Read-X including our methods of evaluating the readability of texts and thematically classifying the texts before concluding with an outline of future work.

Motivation

Recent research has shown that it is possible to identify adult literacy students on the brink of achieving reading fluency in order to provide them with concentrated instruction, dramatically improving their chances of attaining a high quality of life (Strucker et al. 2007). Strucker et al. found that the International Adult Literacy Survey (IALS), a

simple and quick test of literacy skills, can be used to gauge the instructional needs of adult literacy learners. Scores on the IALS range from levels 1 to 5; an IALS score of 3 is correlated with "higher income and less unemployment, increased access to lifelong learning, greater amounts of personal reading for pleasure, and increased civic participation" (Strucker et al. 2007). Those scoring in level 3 are also more likely to have better health and have achieved higher levels of education. Given the pronounced improvement in quality of life found for readers who score in the IALS level 3 compared to the lower levels, it is critical to identify those readers who are just below level 3 and provide concentrated instruction to help them attain at least level 3 literacy.

Low reading proficiency is a widespread problem evident in the performance of adolescents in our schools. The National Center for Education Statistics (NCES) in 2005, the latest year for which data is available, that only 29% of eight graders in the United States achieved proficient or above reading, meaning the remaining 71% of students had only part of the reading skills needed for proficient work at their level or less (Snyder et al. 2006). In 2004 Hasselbring and Goin reported that "as many as 20 percent of 17-year-olds have been estimated to be functionally illiterate, and 44 percent of all high-school students have been described as semi-literate" (Hasselbring & Goin 2004). Reading below grade level is a serious problem for adolescents as it may hinder comprehension of textbooks and classroom materials in all fields. Denti mentions that "most high school textbooks are written at the tenth through twelfth grade levels with some textbooks used for U. S. government written at the seventeenth grade level" (Denti 2004). Seeing how reading skills are tied to academic success, it is no surprise that literacy is linked to later quality of life (Hasselbring & Goin 2004). This fact is especially troubling considering that some demographics are linked to lower achievement in reading (Snow et al. 1998). The fact that children coming from poor families are more likely to have trouble with reading, and that poor readers are more likely to face lower income and higher unemployment, lower health, less civic engagement, points to the critical need for improvements in literacy education (Cole & Hilliard 2006). Weinstein and Walberg studied the factors related to achievement in reading and found that "frequent and extensive engagement in literacy-promoting activities as a young adult was associated with higher scores on literacy outcomes (independent of earlier-fixed characteristics and experiences)," which implies that through ample reading exercise students can achieve literacy regardless of their background (Weinstein & Walberg 1993).

This is our motivation for creating Read-X: with a little help these adults and adolescents can make big advances in reading and in their lives.

Although we have created Read-X with adolescent and adults students in mind, the potential usefulness is not limited to this group. English language learners of all ages have a wealth of specially written ESL literature at their disposal, but authentic reading material is often written at a level beyond their proficiency in English. Authentic reading material is valuable to the students as they prepare to work and read independently in their everyday lives. Read-X can also help ESL/EFL students and teachers locate authentic reading material at appropriate levels of difficulty on the Internet.

Related Work

Studies have been completed on the effectiveness of many existing programs for struggling readers, often specifically for younger readers. Examples of these programs and suggestions for new reading programs for older beginning readers from our review of the literature included features such as increasing motivation to read, making the readings culturally relevant, computer-aided instruction, one-on-one tutoring, and the use of authentic and interesting texts. More empirical research is needed to know how to most effectively help adolescents and adults achieve literacy; Strucker et al. note that such studies are already underway (Strucker et al. 2007).

Just like the old saying, "You can lead a horse to water but you can't make him drink," motivation is very important to helping struggling readers attain literacy. As Guthrie and Wigfield put it, "a person reads a word or comprehends a text not only because she can do it, but because she is motivated to do it" (Guthrie & Wigfield 2000). As they go on to explain, it may be that motivation is what mediates the Matthew Effect, the effect by which, over time, the good readers get better and the poor readers stay poor, because "increasing competence is motivating, and increasing motivation leads to more reading".

Motivating adolescents is especially important. Hasselbring and Goin outlined an ideal program for adolescents, stating that, among other things, it must seem worth doing to them, hide their difficulty from their peers, give them some control over their activities, and provide positive reinforcement instead of more indications of failure (Hasselbring & Goin 2004). To fit these criteria Hasselbring and Goin developed the Peabody Literacy Lab (PLL). In addition to the above criteria, the PLL also feature video anchors for each lesson, which were designed to engage and motivate the students. In a year-long study of at-risk middle and high school students, Hasselbring and Goin found that the PLL helped to improve the reading skills of the 63 students that used the PLL significantly more than the 62 students in the group without it (Hasselbring & Goin 2004).

The simple introduction of technology into the classroom can increase students' motivation. Although this was at a time when computers were more novel in classrooms than they are today, (Kamil et al. 2000) cited the 1985 study by (Lepper & Chabay 1985), who found that "computer-based educational activities can increase factors associated with the intrinsic motivation of students". Kamil et al. concluded that "In the research on the impact of computer use on the classroom structure, the most consistently found effect is an increase in motivation and closely related constructs such as interest and enjoyment of schoolwork, task involvement, persistence, time on task, and retention in school" (Kamil et al. 2000). It is true that computers are much more common today than they ever were, but as they become increasingly commonplace, there is also an increasing quality and quantity of media available on the internet and development of new technologies that may preserve the motivation bump that computers bring to ordinary literacy work (Leu 2000). Surprisingly, Leu found that this increase in engagement when using new technologies was not only on the part of the students, but the teachers as well.

Read-X

Read-X currently runs on computers with Windows XP or Vista and requires an internet connection and Java Runtime Environment (RE) 6. [1] Future versions of Read-X will also be OSX compatible. The current version of Read-X is publicly available from http://net-read.blogspot.com/.

System description

We are developing a web-search tool to locate and evaluate text on the internet. The first release of Read-X, performs the following tasks:

a) searches the web for text related to one more keywords provided by the user,
b) extracts text from the webpages, free of html code,
c) analyzes the readability of each text according to popular readability formulas,
d) classifies the results according to thematic content,
e) returns the readability and thematic classification results as well as the extracted text in editable form.

The user is able to perform a search providing a keyword and optionally selecting desired level of difficulty. The tool processes the search request and analysis in real time and returns the results of the analysis within seconds (assuming a fast connection to the internet). These tasks are described in more detail below.

Internet search

Read-X performs a search of the internet using the Yahoo! Web Services. When the search button is clicked or the enter key depressed after typing in a subject Read-X sends a search request to Yahoo! including the keywords and the number of results to return and receives results including titles and URLs of matching websites in an XML document. The Yahoo! Web Service is freely available for non-commercial use with a limit of 5000 requests per day. If Read-X is deployed for use by a wide number of users, it may be necessary to purchase the ability to process more requests with Yahoo or another search engine.

[1] Most Windows computers have some version of Java installed; users can check if they have the latest version and download it if necessary from http://java.com/en/download/help/testvm.xml .

Text extraction

Read-X then retrieves the html, xml, doc (Microsoft Word document) or PDF document stored at each URL and extracts the human-readable text. The text is extracted from html and xml documents using the scraper provided by Generation Java by Henri Yandell, see www.generationjava.com. The Microsoft Word document scraper is part of the Apache Jakarta project by the Apache Software Foundation, see www.apache.org. The PDF scraper is part of the Apache Lucene project, see www.pdfbox.org. All three of these external tools are available under a common public license as open source software under the condition that any software that makes use of the tools must also make the source code available to users.

Readability analysis

For printed materials, there are a number of readability formulas used to measure the difficulty of a given text; the New Dale-Chall Readability Formula, The Fry Readability Formula, the Gunning-Fog Index, the Automated Readability Index, and the Flesch Kincaid Reading Ease Formula are a few examples. Usually these formulas count the number of syllables, long sentences, or difficult words in randomly selected passages of the text. To automate the process of readability analysis, we chose three Readability algorithms: Lix, Rix, and Coleman-Liau, which were best suited for fast calculation and provide the user with either an approximate grade level for the text or a readability classification of very easy, easy, standard, difficult or very difficult.

When each text is analyzed by Read-X the following statistics are computed: total number of sentences, total number of words, total number of long words (seven or more characters, and total number of letters in the text. Below we describe how each of the three readability scores are calculated using these statistics as well as a proposal for a new readability algorithm that we plan to implement in future versions of Read-X.

***Lix readability formula*:** The Lix readability algorithm distinguishes between five levels of readability: very easy, easy, standard, difficult, or very difficult. If W is the number of words, LW is the number of long words (7 or more characters), and S is the number of sentences, them the Lix index is LIX = W/S + (100 * LW) / W. An index of 0-24 corresponds to a very easy text, 25-34 is easy, 35-44 standard, 45-54 difficult, and 55 or more is considered very difficult.

***Rix readability formula*:** The Rix readability formula consists of the ratio of long words to sentences, where long words are defined as 7 or more characters. The ratio is translated into a grade level using the following table:

Ratio	Grade level
7.2 and above	College
6.2 and above	12
5.3 and above	11
4.5 and above	10
3.7 and above	9
3.0 and above	8
2.4 and above	7
1.8 and above	6
1.3 and above	5
0.8 and above	4
0.5 and above	3
0.2 and above	2
Below 0.2	1

Table 1: Long words to sentences ratio translation to grade level using the Rix readability formula.

***Coleman-Liau readability formula*:** The Coleman-Liau readability formula is similar to the Rix formula in that it gives the approximate grade level of the text. Unlike the Lix and Rix formulas, the Coleman-Liau formula requires the random selection of a 100 word excerpt from the text. Before the grade level can be calculated, the cloze percent must be estimated for this selection. The cloze percent is the percent of words that, if deleted from the text, can be correctly filled in by a college undergraduate. If L is the number of letters in the 100 word sample and S is the number of sentences, then the estimated cloze percent is C = 141.8491 – 0.214590 * L + 1.079812 * S. The grade level can be calculated using the Coleman-Liau formula, where grade level is -27.4004 * C + 23.06395. In the Read-X display we round the final result to the nearest whole grade level.

***A new readability algorithm in progress*:** In addition to the automation of text analysis we are developing a new readability algorithm for automated analysis of text that can be customized for each student who uses Read-X. Our new algorithm will take into account the readers content knowledge when evaluating the readability of a text. Simple readability algorithms like Lix, Rix and Coleman-Liau use word length as an approximation for difficulty; words with over seven characters are considered long and therefore more difficult than shorter words. This makes the algorithms easy to use and automate, but longer words aren't always more difficult than shorter ones. For example, a student who is knowledgeable or likes to read about music would probably be familiar with words like harmony, acoustic, bluegrass, and instrument and not necessarily words like alfalfa, garrison, or processor, which are not common in music texts. The lengths of these words are about the same, but the first group of words should contribute to a lower overall readability score than the second. For this reason, we are currently developing new readability formulas sensitive to the user's prior reading experience. As a first step towards this direction we have built corpora from different thematic areas and computed vocabulary frequencies per area.

Thematic classification

In addition to readability analysis, each text is thematically classified to help the user better gauge the content of the text. For example, if the user enters the keywords "black eyed peas" they could end up with texts about cooking or about the popular band of the same name; the thematic classification can help separate the music texts from the texts about food. In the next release of Read-X, the user will also be able to pre-filter the search by selecting the thematic area of their choice from the main menu.

For the thematic classification of the extracted text we use machine learning methods. Specifically, we use the Mallet classifier (Machine Learning for Language Toolkit).[2] The Mallet classifier is trained on a three million corpus that we have built for this purpose. The corpus has been hand-selected and tagged for thematic area. The tags correspond to eight themes: Arts, careers and business, literature, philosophy and religion, science, social studies, sports and health, and technology. The classifier is fully integrated in Read-X and is able to thematically classify texts in real time and also give a percent value representing how well the text fits within the identified category.

Displaying results

Below is a screenshot of Read-X. At the very top we have entered "magnetism" as a keyword and clicked the "Search" button or pressed enter to begin the search. The small progress window keeps track of the overall progress of the search while results appear one at a time in the scrollable table.

At the top left of the screen there are two menus. The first, labeled "Menu", provides access to help and information about Read-X. The second menu, labeled "Settings", allows the user to select the number of results to return and also to identify which if any subjects the user is most familiar with. The last feature is intended for future use with our new readability formula and is not integrated with the readability analysis in this version of Read-X. On the rightmost column, the user can click on 'view text' to view, edit and save the text in a word processing editor that pops in a separate window.

[2] Mallet has been developed at the University of Massachusetts Amherst Computer Science Department, see www.cs.umass.edu/~mccallum/mallet. Mallet is available for use under the Common Public License, see www.opensource.org.

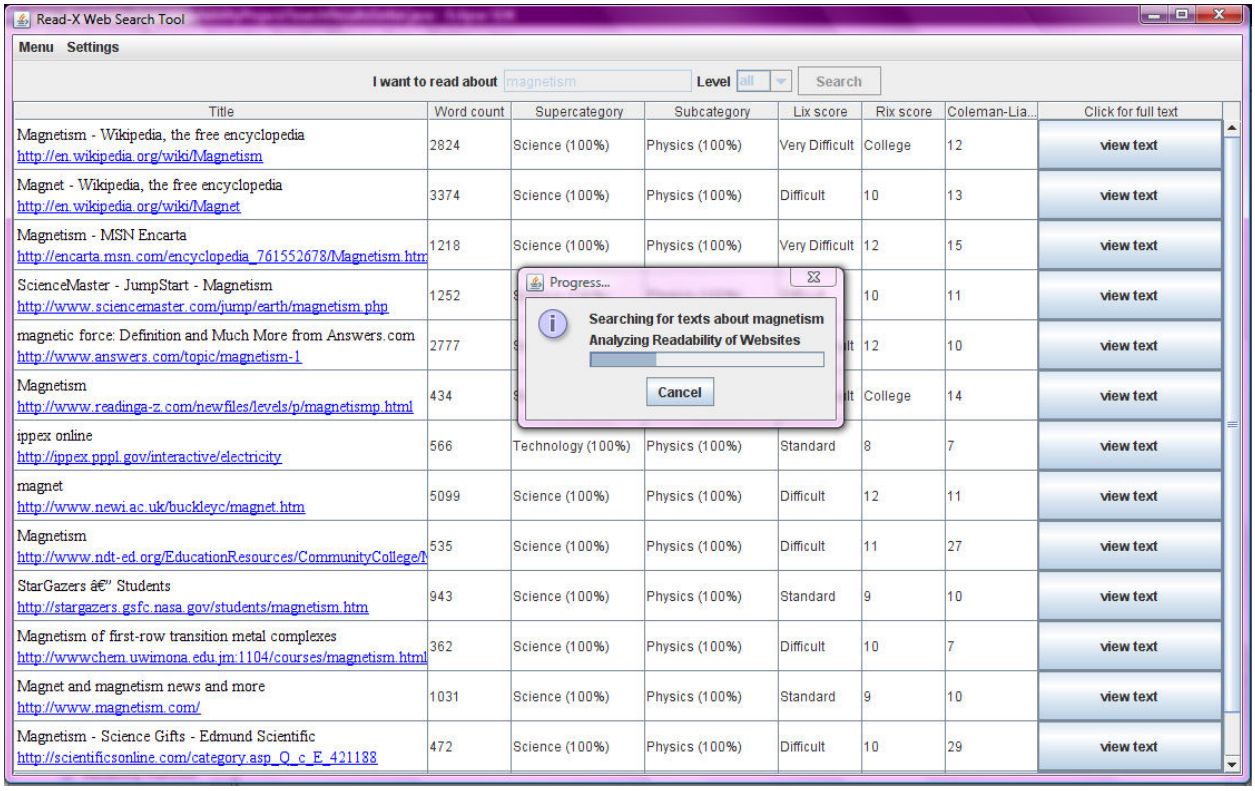

Figure 1: Screenshot of Read-X. For each text the user is given a link to the text on the internet, the word count, thematic classification, and three measures of readability using the Lix, Rix and Coleman-Liau algorithms. In the last column is a button to open a text-editor window where the text can be viewed, saved, printed or edited without HTML or other web content present.

Conclusion and Future Work

As there is great need for appropriately challenging reading material for mature, low-level readers, much work will build from this first project. Read-X is available for further user testing and peer review, the findings of which will be monitored and improvements made as necessary. A new feature that might be added to the web search tool at a later date is support for different languages. Readability formulas for foreign languages already exist, and some of the formulas described above can be used to measure readability in western European languages as accurately as in English. In this way, the web-search tool and readability assessment method would benefit not only American students of foreign languages, but also educators all over the world.

Extensive work in psycholinguistics has investigated the effect different syntactic structures might have in slowing down reading times. For example, Gail McKoon and Roger Ratcliff studied the role that syntax and sentence structure play in reading comprehension, specifically with regard to the ambiguity and altered meaning of reduced sentences. For example, "*The hangman executed by the government* does not convey the same thing as *The hangman who was executed by the government*" (McKoon & Ratcliff 2007). McKoon and Ratcliff found that the longer phrase was actually more comprehensible. The results of this research could be used to improve the subtlety of measurement of existing readability assessment formulas, which employ superficial measures such as counting syllables, words or sentence length. It is computationally feasible to perform such syntactic analyses of digitized texts, but these require significant processing time. Before these sophisticated evaluations of text comprehensibility can be implemented in a web-based search tool, methods must be developed for improving the speed of processing in real-time.

For the next version of Read-X our plan is to have each user create a profile and indicate his or her favorite subjects, or subjects in which the he or she is very knowledgeable. This information will help the application to more

accurately estimate the difficulty of each text for the student using the new readability formula that we plan to develop as described above. We have calculated word frequencies for each thematic area for the texts in our three million word corpus. For each of the eight thematic categories, arts, careers and business, literature, philosophy and religion, science, social studies, sports and health, and technology, we can use the word frequency lists to identify words that the user is likely to know, and use that information to alter the readability analysis of texts for students based on their interests and existing knowledge. In the future, the system will also be able to track the texts that the user reads and the new vocabulary she has been exposed to in order to update the algorithm over time.

Our ultimate goal is to make Read-X available as a web-based application that students and teachers can use instead of other search engines to locate reading material online.

Acknowledgments

Funding for the development of Read-X during the summer of 2007 was provided by the Graduate and Professional Student Alliance (GAPSA) and the Office of the Provost of the University of Pennsylvania through the 2007 GAPSA-Provost Award for Interdisciplinary Innovation.

References

Cole, Juanita McLean, & Hilliard, Veleshia Rhonda (2006). The Effects of Web-Based Reading Curriculum on Children's Reading Performance and Motivation. *Journal of Educational Computing Research* 34 (4): 353-380.

Denti, Lou (2004). Introduction: Pointing the Way: Teaching Reading to Struggling Readers at the Secondary Level. *Reading and Writing Quarterly* 20: 109-112.

Guthrie, John T., & Wigfield, Allan (2000). Engagement and Motivation in Reading. *Handbook of Reading Research* 3. Michael L. Kamil, Peter B. Mosenthal, P. David Pearson, and Rebecca Barr, Eds. Lawrence Erlbaum Associates, New Jersey: 403-422.

Hasselbring, Ted S., & Goin, Laura I. (2004). Literacy Instruction for Older Struggling Readers: What is the Role of Technology? *Reading and Writing Quarterly* 20: 123-144.

Kamil, Michael L., Intrator, Sam M., & Kim, Helen S. (2000). The Effects of Other Technologies on Literacy and Literacy Learning. *Handbook of Reading Research* 3. Michael L. Kamil, Peter B. Mosenthal, P. David Pearson, and Rebecca Barr, Eds. Lawrence Erlbaum Associates, New Jersey: 771-788.

Leu Jr., Donald J. (2000). Literacy and Technology: Deictic Consequences for Literacy Education in an Information Age. *Handbook of Reading Research* 3. Michael L. Kamil, Peter B. Mosenthal, P. David Pearson, and Rebecca Barr, Eds. Lawrence Erlbaum Associates, New Jersey: 743-770.

McKoon, G., & Ratcliff, R. 2007. Interactions of meaning and syntax: Implications for Models of Sentence Comprehension. *Journal of Memory and Language* 56: 270-290.

Morgan, Paul L., & Fuchs, Douglas (2007). Is there a bidirectional relationship between children's reading skills and reading motivation? *Exceptional children* 73 (2): 165-183.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). Committee on the Prevention of Reading Difficulties in Young Children, National Research Council: *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Snyder, T.D., Tan, A.G., & Hoffman, C.M. (2006). *Digest of Education Statistics 2005* (NCES 2006-030). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Strucker, John, Kentaro Yamamoto, & Kirsch, Irwin (2007). The Relationship of the Component Skills of Reading to IALS Performance: Tipping Points and Five Classes of Adult Literacy Learners. *NCSALL Reports* #29. Boston: National Center for the Study of Adult Learning and Literacy (NCSALL).

U.S. Department of Education (2005). Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Reading Assessment. *http://nces.ed.gov/nationsreportcard/nde/* Accessed April 11, 2007.

Weinstein, Thomas, & Walberg, Herbert J. (1993). Practical literacy of young adults: educational antecedents and influences. *Journal of Research in Reading* 16 (1): 3–19.