

Automatic Readability Assessment for People with Intellectual Disabilities

Lijun Feng

City University of New York

lijun7.feng@gmail.com

Abstract

My research goal is to advance our understanding of, and quantify, what makes a text easy or difficult to read, in particular for readers with intellectual disabilities. Previous research in automatic readability assessment has looked at a limited class of lexical and syntactic properties of texts. Moreover, these models are usually not targeted towards any particular group of readers. In my own work, by contrast, I have used sophisticated computational tools to build an automatic readability metric that exploits global semantic (discourse level) properties of a text, in addition to well-studied lexical and syntactic features. Our preliminary results (Feng et al., 2009) confirm the value of discourse attributes. My research is targeted towards understanding the particular difficulties faced by readers with intellectual disabilities. The ultimate goal is not simply to model and understand readability issues, but also to aide in the development of automatic language processing tools that can rewrite texts to be more readable.

Motivation and Background

According to the 2006 American Community Survey (U.S. Census Bureau, 2006), about 5% of the civilian noninstitutionalized population, approximately 13.5 million people age 16 or above in the United States have mental disabilities, with intelligence test scores of 70 or below. Among this group of people, about 85% are in the category of mild mental retardation (IQ range 50–75) (Drew and Hardman, 2004). We will use the term “intellectual disabilities” (ID) or “mild intellectual disabilities” (MID) henceforth. People with ID face many challenges in their daily lives; one of these challenges lies in the area of reading literacy. A study conducted by Jones et al. (2006) assessing the reading comprehension of adults with MID reported that the average reading skills of subjects were below that of average 7-year-old readers without disabilities.

Several factors contribute to the lower literacy skills of adults with ID. Above all, the limitation of their cognitive functioning affects their reading comprehension directly. Research has shown that people with ID are often better at sounding words out rather than comprehending their meaning (Drew and Hardman, 2004). It has also been shown that they are often slow at resolving the identity of proper names and encoding them into a symbolic form in their memory (Hickson-Bilsky, 1985). These individuals often have problems remembering and inferring information from text because of their limited working memory (Fowler, 1998). Consequently, as part of the previously read units are lost from the working memory, they often have difficulty with integrating complex information into a cohesive semantic representation (Hickson-Bilsky, 1985), which likely results in their lower reading comprehension.

It is difficult to find reading materials for individuals with MID that are (1) of interest to them and (2) at the right reading level. Reading materials at lower reading levels are typically

written for children, and texts written for adults without disabilities often require a high level of linguistic skills and sufficient real world knowledge, which these individuals often lack. The lack of appropriate reading materials may also discourage adults with ID from practicing reading, thus diminishing their already low literacy skills.

The need to identify or reformulate texts suitable for lower reading levels is not unique to people with ID. Children, second language learners, and adults with low literacy skills can also benefit from such texts. However, manually adapting written texts is both time and labor intensive. In the past decade, natural language processing (NLP) techniques have been used to develop automatic text simplification systems to assist human readers (Devlin, 1999; Inui et al., 2003). Research has focused mainly on lexical and syntactic simplification. Lexical simplification often uses word frequency or predefined wordlists to identify difficult words and replaces them with simpler synonyms. Syntactic simplification often uses dependency-tree structures and pattern recognition techniques to identify complex syntactic constructs, such as relative clauses, passive voice, and conjoined sentences; transformation rules are then applied to change these constructs into shorter or plainer sentences.

People with MID would certainly benefit from texts simplified in this fashion. However, synonym-replacement and syntax-tree simplification alone cannot fully cover the needs of this group of users, because, in addition to challenges that come from lexical and syntactic factors, they have other difficulties with processing written information as discussed above. Moreover, most earlier text simplification systems process input text one sentence at a time, which inevitably results in increased length of the simplified document, because long and complex sentences are often split into multiple shorter sentences. The resulting increased length of the whole document can pose another challenge to readers with MID because it requires processing and storing more information.

In order to meet the special needs of this group of underrepresented individuals, we are ultimately interested in designing and implementing an automatic text simplification system that modifies a text at the discourse level, in addition to lexical and syntactic simplification. This entails high-level semantic simplification, whereby the most relevant information is retained and less relevant information simplified or completely left out (Feng, 2008)

The following proposed dissertation research is the first step of a longer-term research project into a discourse-level text simplification system. We face several open, foundational questions, which are both self-contained and crucial for further research, and thus form a stand-alone dissertation project. There are two major research questions that are at the center of the design and implementation of such a text simplification system (Inui et al., 2003): (1) How do we identify which portions of a text will pose difficulty for our users? (2) When there are several possible simplification choices, how do we decide which is the optimal one to choose for our users? We need a reliable reading assessment tool to guide the system's actions during the simplification process and to evaluate the difficulty posed by a given text, whether found, constructed by experts, or simplified automatically.

Relevant Literature and Previous Work

Extensive research has been conducted in the past 80 years to understand what affects the readability of a text and how to assess its reading difficulty. To make it easier for people to judge the reading difficulty of a text, grade levels or number of years of education

required to completely understand a text are commonly used as index for reading difficulty.

Many traditional readability metrics use simple linear functions with two or three shallow language features to model the readability of a given text. For example, the Flesch-Kincaid grade level formula (Flesch, 1979) uses average sentence length and average syllables per word to calculate the grade level of a text. Similarly, the Gunning FOG index (Gunning, 1952) uses average sentence length and the percentage of words with at least three syllable as parameters. These traditional metrics are widely used, especially in educational settings, partly because they are easy to calculate. However, these metrics do not always capture the reading complexity of a text accurately, which has been confirmed by several recent work in the field (Si and Callan, 2001; Feng et al., 2009; Petersen and Ostendorf, 2009). For example, the preliminary results of our research show that, tested on 1433 articles labeled with grade levels ranging from 2 to 5, Flesch-Kincaid metric only predicted 29 (2%) of them with correct grade level, while our metric predicted 881 (61%) correctly.

It is understandable why traditional readability metrics are not reliable, because the complexity of a writing lies in much more complicated factors than just average sentence length or average syllables per word. Moreover, short sentence length does not necessarily indicate easy readability (consider poems, for instance), and there are many infrequently used words with only a few syllables.

Recent work on readability deployed sophisticated natural language processing techniques, such as parsing and statistical language modeling, to capture more complex linguistic features and used statistical machine learning tools to build readability metrics. Si and Callan (2001) used unigram language models to capture content information from scientific web pages. Combined with surface linguistic features, they built a classifier with a linear model to predict the reading difficulty of these web pages with success. Collins-Thompson and Callan (2004) adopted similar approach and used a smoothed unigram model to predict the grade levels of short passages and web documents. Michael J. Heilman and Eskenazi (2007) continued using language modeling to predict readability for first language texts, furthermore, they extracted grammatical features from parsed documents to build a classifier for readability prediction for second language texts. Schwarm and Ostendorf (2005) used support vector machines to combine features from traditional reading level measures, statistical language models and automatic parsers to assess reading levels.

Recent work on readability all confirms the benefits of using statistical language models and/or parsers over traditional measures. It has also be shown that, although some features may outperform the others, when they are used combined, they often help improve the overall performance of the readability metrics. Our preliminary results are consistent with these findings.

Methodology and Data

Our research goal is to build and evaluate an automatic readability assessment tool that accurately models the reading difficulty of a text for people with intellectual disabilities.

The problem with building such a metric lies in that text difficulty is not directly observable, we need to do user studies and/or find proxy variables that associate with reading difficulty. Our general methodology relies on the following five proxies. (1) Paired

original/simplified texts. A common assumption is that simplified texts should be easier to read. Paired texts provide valuable clues on how texts with identical subject matter differ. During our modeling, we will use paired texts to analyze and select features that distinguish the simplified texts most from the original ones. (2) Grade levels. They are a commonly accepted index for reading difficulty of a text, especially in educational settings. Reading difficulty increases with grade level. It is an open research question in our study whether grade levels are appropriate readability indicators for our users. Our hypothesis is that text properties that influence reading difficulty for average readers are qualitatively (but perhaps not quantitatively) the same for readers with ID. This hypothesis says that what is hard for average readers is also hard, perhaps to a different degree, for readers with ID. Put yet differently, we are not aware of any text properties that cause problems only for readers with ID. Vice versa, we think a readability metric that is designed for a specific group of users can be useful for the general public as well. Another reason to look at grade levels is because they have been widely used in the existing literature. Novel readability metrics should try to predict grade levels so that they may be compared with existing approaches. (3) Subjective ratings by experts. We will ask experts who specialize in working with adults with ID to rate text difficulty. The motivation behind this is as follows: (a) We rely on their expertise to help us identify factors that may play an important role in affecting reading difficulty for our users. (b) Subjective expert ratings are much more reliable and easier to obtain than from target users. We will evaluate subjective ratings by checking inter-rater agreement, as well as correlation with grade levels and subject ratings and observations. (4) Subjective (introspective) ratings by users. This will probably be especially problematic in our study, as the users' subjective judgment may not be fully reliable because of their cognitive impairments. Many research questions remain open as how to design and conduct studies with adults with ID to get effective and valid feedback. However, we believe direct user feedback is valuable in our user-specific study. (5) Objective observations in user studies. We will present our target users with texts at a variety of difficulty levels and record their reading times. Subject will further answer simple comprehension questions after reading, and we will analyze the accuracy of their answers. This will give us the most direct clues about the difficulties faced by our target user group, even though we will need to account for per-subject and other effects. For developing our readability metric, we want to combine all the above observations to get at those underlying text properties that are associated with reading difficulties.

We plan to develop three groups of features based on the analysis of the paired original/simplified texts: (1) traditional shallow features, (2) syntactic features, and (3) novel discourse features. Our shallow and syntactic features (1) are mainly inspired by previous work in the field. Shallow features include those that are often used by traditional readability metrics, such as average sentence length per word, average number of syllables per word, total number of polysyllabic (3) words, percentage of polysyllabic words. Our syntactic features (2) include (a) parse tree features, such as average parse tree height, ratio of terminal and non-terminal nodes, number of relative clauses, noun phrases, verb phrases and prepositional phrases, etc.; (b) cross-domain perplexity features captured by various statistical language models; and (c) part-of-speech features, such as average number of adjective, adverbs, past participles, past tense verbs, present participles, modal verbs, and infinitive markers.

Our selection of discourse features (3) will focus on those that are cognitively motivated by our user characteristics. Our hypothesis is that, aside from infrequently used word and

complex syntactic constructs that even average readers find difficult, the density of the amount of information introduced in a text and whether the information needs to be inferred may pose particular challenges for our users. Named entities, such as people's names, locations, organizations, together with general nouns, serve as major information carriers in a text. We hypothesize that the more entities there are to be processed, the more they will overwhelm our users' already slow semantic decoding and limited working memory. Similarly, at a even higher discourse level, factors such as the amount of topics introduced in a text, where each of them starts and ends, where some of them overlap, require excessive working memory to process and keep track of. More-over, our target users often have problems remembering and inferring information from a text (Fowler, 1998), so information that is not stated directly and needs to be inferred by readers can be challenging for our users. Therefore, our discourse features will include (a) entity density features, such as average number of unique entities (people, location, organization and general nouns) per sentence, total entity mentions per sentence, or the total counts of both in a text; (b) discourse topic features, such as the total number of topics, the average span of a topic, number of overlapping active topics, etc; and (c) coreference features, such as average distance between pronouns and their referents, and the number of nouns and/or pronouns that refer to the same object or person.

We use various existing and novel NLP techniques and toolkits to extract features. To extract syntactic features, we use the well known Charniak parser (Charniak, 2000) for automatic syntactic analysis. We use LingPipe (<http://alias-i.com/lingpipe>) to extract named entities and solve coreference related issues. We use the lexical chain software developed by Galley and McKeown (2003) to annotate synonyms, hypernyms and hyponyms. The SRI Language Modeling Toolkit (<http://www.speech.sri.com/projects/srilm>) and the CMU-Cambridge Statistical Language Modeling Toolkit (<http://www.speech.cs.cmu.edu/SLM>) are used to train various language models.

To investigate the possible impacts of these features on readability and to construct our metric, we use statistical and machine learning techniques, in particular support vector machines (SVMs), to train various classification and regression models on labeled data. Our preliminary experiments use LIBSVM (Chang and Lin, 2001), an integrated software for support vector classification and regression, which supports efficient multi-class classification and provides an automatic tool to perform grid search for optimal parameters, which are essential in improving the performance of our models (C.-W. Hsu, 2003).

An ideal corpus for our research would be a set of original texts paired with their simplified versions adapted just for our target users with MID. Unfortunately, there does not exist such a corpus yet, to the best of our knowledge. In order to develop our methods and test our hypothesis, we collected two comparable corpora from two sources: literacynet.org and Encyclopedia Britannica (Barzilay and Elhadad, 2003). Both corpora consists of paired original/simplified texts. We collected a third corpus from the Weekly Reader (<http://www.weeklyreader.com>), an online publisher producing magazines for students in grades pre-K to 12. This corpus consists of 1433 articles in total with grade levels labeled ranging from grade 2 to 5. We use this corpus to train various classification and regression models using SVM, and we test the baseline performance of our models on the same data and compare them with existing literature. At the same time we will create our own user-

specific paired corpus for user studies. We will collect texts that are of interest to our users and recruit experts on adults with ID to simplify them.

Preliminary Research

Currently, we have collected two unlabeled paired corpora (LiteracyNet and Encyclopedia Britannica) and a labeled corpus (Weekly Reader). We have developed 49 features in total and implemented a baseline library to calculate them. Using these features and the labeled data we have built our models and evaluated them.

Our preliminary results show that our discourse features are competitive to shallow and syntactic features, they help improve the overall performance of our metric when used combined (see Table 1). Compared with previous work, our novel readability metric are very promising in predicting correct grade levels. It outperformed the traditional Flesch-Kincaid readability metric by remarkable margin (see Table 2A), and remains competitive with other methods similar to our approach. Compared with a similar study conducted by Petersen and Ostendorf (2009), our metric outperformed theirs with high margin on articles with lower grade levels (see Table 2B and 2C). For further details, see our forthcoming paper (Feng et al., 2009), accepted for publication at EACL-09.

	A. Shallow features			B. Syntactic features			C. Discourse features		
	Prec	Recall	F-score	Prec	Recall	F-score	Prec	Recall	F-score
Gr 2	0.51	0.49	0.50	0.64	0.55	0.59	0.56	0.47	0.51
Gr 3	0.49	0.54	0.51	0.52	0.56	0.54	0.47	0.54	0.50
Gr 4	0.46	0.43	0.45	0.48	0.49	0.48	0.53	0.44	0.48
Gr 5	0.64	0.64	0.64	0.66	0.65	0.66	0.66	0.73	0.69

Table 1: Comparison of precision, recall, and F-score for SVM trained with shallow features, syntactic features and discourse features on the Weekly Reader data (10-fold cross-validation).

	A. Flesch-Kincaid			B. Peterson et al 2009			C. Our metric (10-f CV)		
	Prec	Recall	F-score	Prec	Recall	F-score	Prec	Recall	F-score
Gr 2	0.01	0.03	0.02	0.38	0.61	0.47	0.70	0.58	0.63
Gr 3	0.01	0.00	0.00	0.38	0.87	0.53	0.55	0.58	0.56
Gr 4	0.19	0.03	0.05	0.70	0.60	0.65	0.54	0.52	0.53
Gr 5	0.83	0.02	0.04	0.75	0.79	0.77	0.68	0.72	0.70

Table 2: Comparison of Flesch-Kincaid, Peterson et al's recent work and our metric based on 10-fold cross-validation

Future Work

In the next stage of this research, we will continue development of more discourse features using coreference resolution software and additional complex syntactic features and incorporate these new features into our metric. At the same time we will recruit experts to simplify texts for our users, ask for experts' subjective ratings, recruit adults with MID as test subjects, design questionnaires for user studies, and conduct user experiments. In Fall 2007, we conducted a pilot study with 14 adults with ID, which provided us with valuable feedback on many aspects, such as what kind of objective measures worked, what still needs to be

improved, and what did not work. We also learned how to interact with our test subjects during the study to obtain qualitative feedback.

In the end, we will use our metric to predict readability of the texts that are presented to the experts and the test subjects. We will analyze the correlations of the predictions by our metric with experts' subjective ratings and user feedbacks.

References

1. Chih-Jen Lin C.-W. Hsu, Chih-Chung Chang. 2003. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.
2. Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: a Library for Support Vector Machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
3. Eugene Charniak. 2000. A maximum-entropy-inspired parser. In Proceedings of the 1st Conference of the North American Chapter of the ACL, page 132139.
4. Kevyn Collins-Thompson and Jamie Callan. 2004. A language modeling approach to predicting reading difficulty. In In Proceeding of HLT/NAACL.
5. Siobhan Devlin. 1999. Simplifying natural language for aphasic readers. Ph.D. thesis, University of Sunderland, UK.
6. Clifford J. Drew and Michael L. Hardman. 2004. Mental retardation: A lifespan approach to people with intellectual disabilities. Merrill, Columbus, OH.
7. Lijun Feng. 2008. Text simplification for people with intellectual disabilities. In 10th ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '08). Halifax, Nova Scotia, Canada. Doctoral Consortium.
8. Lijun Feng, Noemie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In The 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09). To appear in EACL, Athens, Greece.
9. Rudolf Flesch. 1979. How to write plain English. Harper and Brothers, New York.
10. Anne E. Fowler. 1998. Language in mental retardation. In Handbook of mental retardation and development, pages 290–333. Cambridge University Press.
11. Michel Galley and Kathleen McKeown. 2003. Improving word sense disambiguation in lexical chaining. In Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003).
12. Robert Gunning. 1952. The Technique of Clear Writing. McGraw-Hill.
13. Linda Hickson-Bilsky. 1985. Comprehension and mental retardation. International Review of Research in Mental Retardation, 13:215–246.
14. Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, and Ryu Iida. 2003. Text simplification for reading assistance: A project note. In Proceedings of the Second International Workshop on Paraphrasing, pages 9–16.
15. W. Jones, K. Long, and W. M. L. Finlay. 2006. Assessing the reading comprehension of adults with learning disabilities. Journal of Intellectual Disability Research, 50:410–418.
16. Jamie Callan Michael J. Heilman, Kevyn Collins-Thompson and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL'07),, pages 460–467.
17. Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. Computer Speech and Language, 23:89–106.

18. Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In Proceedings of ACL.
19. Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In Proceedings of the tenth international conference on Information and knowledge management.
20. U.S. Census Bureau. 2006. American community survey/puerto rico community survey 2006 subject definitions. Available online at http://www.census.gov/acs/www/Downloads/2006/usedata/Subject_Definitions.pdf.

About the author:



Lijun Feng is a PhD candidate in the Computer Science Department at the Graduate Center of the City University of New York (CUNY), where she is working with Prof. Matt Huenerfauth, her thesis advisor. Her research interests include natural language processing (NLP), in particular readability, text simplification, and text comprehension. Her thesis research combines NLP and machine learning techniques to build and evaluate an automatic text readability assessment tool, with special focus on readers with intellectual disabilities. Before entering CUNY, she obtained my masters degree in Computer Science from Brooklyn College in May 2005.