Readability of Linked Data with Experiments on School Textbooks

Chandni Goplani[1], Prof. Nidhi Madia[2]
[1, 2] SOCET, Ahmedabad

Abstract- *This paper describes ongoing experiments to check the readability of linked documents. A user study was conducted on school textbooks by manually adding hyperlinks to some of the central concepts in the text. The study was conducted on school children whose second language was English. It was found statistically significant over plain texts. We also describe our experiment to automatically add hyperlinks to text. Other than adding hyperlinks we also did some text simplification to understand the effect on comprehensibility. Finally it discusses a possible way to measure that change in readability due to linking. A formula is proposed for readability of linked documents that gives a reasonable approximation to the readability.*

Keywords- Readability, Readability Formula, Labeling Contents, Text Simplification

I. INTRODUCTION

Readability of a text generally refers to how well a reader is able to comprehend the content of a text, through reading (Dale and Chall, 1948). Readability formulas quantify readability. Though these formulas have limitations(Redish, 1981; Drury, 1985; Bruce et al., 1981), they are an attractive prospect used by many authors and publishers to evaluate and revise their text (Bruce et al., 1981)[1]. Well known readability formulae for English language are the Flesch Formulae, the Dale-Chall Formula (DuBay, 2007), the Gunning Formula(Gunning, 1952) [16], the SMOG formula (McLaughlin, 1969), the Fry Formula (Fry,1968) and a few others [2]. In text simplification we try different methods to simplify the text and improve its readability, that is, obtain a better readability score on one of the above mentioned readability formulas. The research in Automatic Text Simplification is accurately surveyed by Shardlow (2014) [11]. Text enrichment with links to a knowledge base can also improve its comprehensibility and hence readability. Since the readability formulas mentioned here work only for an individual piece of text, they need to be modified to obtain an objective approximation to the readability of linked data. We experiment with these notions and report out findings in this paper.

II. LINKED DATA AND TEXT ENRICHMENT

There is a huge amount of information on the internet. The information is in a hyperlinked format and hence navigating knowledge spaces is often required (West and Leskovec, 2012). The links are used when people want additional information. The additional information may improve the understanding and hence tends to improve readability also. An approach to provide links to enhance readability has already been used for documents of programming interfaces by enriching them with examples (Kim et al. , 2013) [10]. Similar approach can be used for normal text also. Our goal is to study the impact of providing links to the text in school textbooks and hence verify if it can improve readability specifically for school children.

III. EXPERIMENT

The experiment to study the impact of linked data on readability of school textbooks was conducted online in a supervised environment in a school computer laboratory.

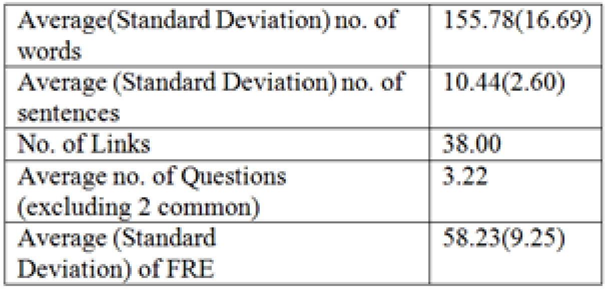
For the experiment nine passages were selected randomly from the science textbooks of class 9. A subject expert was asked to pick up the words or phrases from the passages for which links could be provided. He chose the words or phrases which, according to him, were difficult for the students or for which extra information could be helpful. On the basis of those words or phrases relevant passages or portions of articles were chosen from Wikipedia. These pieces of texts were linked to the original textbook passages. Comprehension questions were chosen for each passage used in the experiment. With each of the passages there were the following two common questions asking the participant for a score on Likert Scale (Likert,1932)

1. Rate the readability (how much you understand) of the passage. Give number (decimal) in range 1-10 (1- very difficult and 10- very easy).
2. How interesting did you find the paragraph? Give number (decimal) in range 1-10 (1- very uninteresting and 10- very interesting).

Similar questions were used by Sinha et al.(2014).Flesch Reading Ease (FRE) was used to measure the readability of the passages. Higher the FREscore, better the readability. The statistics for the passages used are as given in

1. The average Flesch Reading Ease for the links was 35.78 and the average number of words per link was 53.10.

Table 1: Paragraph Analysis



Average(Standard Deviation) no. of words	155.78(16.69)
Average (Standard Deviation) no. of sentences	10.44(2.60)
No. of Links	38.00
Average no. of Questions (excluding 2 common)	3.22
Average (Standard Deviation) of FRE	58.23(9.25)

The passages were divided in 6 sets and 3 groups. One set in each group contained 3 passages with links and the other set in the same group contained the same 3 passages without links. The experiment was conducted in three sessions. Class 8 students volunteered as participants. In each session the participants were divided into two groups. One group was assigned set with linked paragraphs and the other group was assigned set without linked paragraphs. The students were given instructions on what they had to do before logging in to the website. They were told that they were supposed to give an online comprehension test. They were assigned their respective sets of three paragraphs. They were instructed to read a paragraph and answer questions following it one by one. One of the two groups was told about the links and instructed to use them if they required additional information. All students were explained about the Likert scale questions at the end of each paragraph and asked to answer them on the basis of their personal experience. They were instructed to submit the test after informing the supervisor. The entire experiment was closely supervised and timed.

3.1 Results

The experiment to study the impact of linked data on readability of school textbooks was conducted online in a supervised environment in a school computer laboratory.
The total number of participants in the experiment was 65 (23 boys,42 girls). Out of these, 12 boys attempted sets with links while 11 attempted the ones without links. From the girls 28 attempted sets with links and 14 attempted the ones without links. Hence a total of 40 participants attempted sets with links and rest 25 without links.

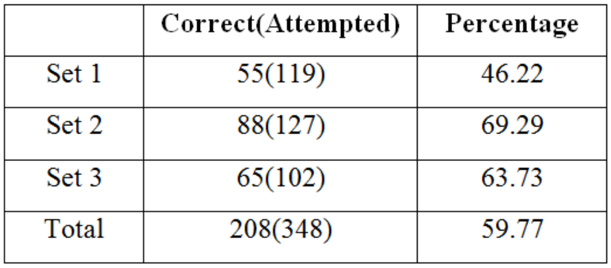
95 passages with links and 80 without links were attempted. The average time taken per passage with links was about 9 minutes 32 seconds while for a passage without links it was 7 minutes 16 seconds per passage.

Out of total 40 participants who were assigned sets with links, the links were used by only 23 of the participants and that too in a very small number. Only 16.5 percent of the total possible links were used.

Regarding the comprehension questions out of all attempted questions, not considering the Likert scale questions, the statistics obtained are as given in 2 and 3 for sets with and without links respectively.

The average readability score given by participants who were assigned sets with links was 9.15 with a standard deviation of 1.5. The average interest score given by them was 8.73 with a standard deviation of 2.02. The average readability score given by participants who were assigned sets without links was 8.96 with a standard deviation of 1.87. The average interest score given by them was 8.77 with a standard deviation of 2.22.

Table 2: Student Performance in sets with links

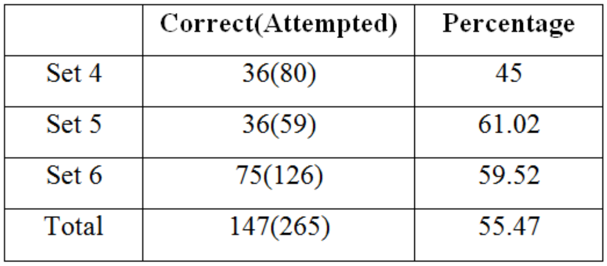


	Correct(Attempted)	Percentage
Set 1	55(119)	46.22
Set 2	88(127)	69.29
Set 3	65(102)	63.73
Total	208(348)	59.77

3.2 Observations from User Study

From the results obtained, it is seen that a very low number of available links were used by the students. The readability score and the interest score given by both, participants assigned set with links and those without links, are comparable.

Table 3: Student Performance in sets without links



	Correct(Attempted)	Percentage
Set 4	36(80)	45
Set 5	36(59)	61.02
Set 6	75(126)	59.52
Total	147(265)	55.47

The student’s assigned sets with links outperform the other students by a fair amount in terms of comprehension questions. A two tailed T-test performed on the no. of correct answers given by students in both the samples assuming unequal variance gives a p-value of 0.024. This gives a significant backing up to the assumption that links do in fact enhance readability. Here we observed that though the average FRE scores for the links is less than that of the passages the overall comprehensibility improves when links are used. This encourages us to the hypothesize that if we apply text simplification to both the passages and the links and then check for readability it should be perform better than just with links. Hence we propose to combine two approaches: Text simplification and Text enrichment by links to a knowledge base to improve the readability of a text.

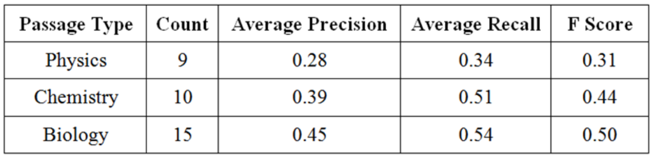
IV. IMPACT OF TEXT SIMPLIFICATION

If we want to combine the two approaches mentioned in previous section for improving readability and automate the two processes, we divide it into two tasks: Automatically selecting words to be linked and text simplification of passage and links.

For the first task we used two methods. In the first method we calculated the term frequency for each term in all science textbooks from class 6 to 12. Each chapter was considered one document. As the rarely occurring terms are expected to have more information we multiplied the inverse of term frequency to the inverse of the number of documents in which the term occurs and prepared a rank list of all terms. Based on a low threshold the terms to be linked were selected. The method was applied to 34 passages, including the 9 passages used in the user study and results were compared to manually tagged passages. The results are given in 4.

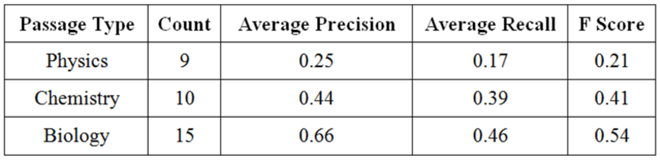
In the second method we used the rank list of Kucera Francis Frequency(Ku et al. , 1967; Quinlan, 1992) and use a threshold 0 to decide the words to be linked. The results are as shown in 5.

Table 4: Results by ITF – IDF



Passage Type	Count	Average Precision	Average Recall	F Score
Physics	9	0.28	0.34	0.31
Chemistry	10	0.39	0.51	0.44
Biology	15	0.45	0.54	0.50

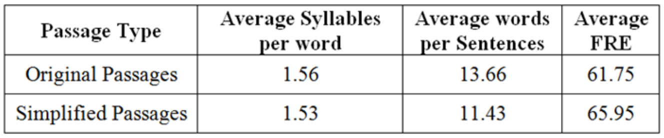
Table 5: Results by Kucera Francis Frequency



Passage Type	Count	Average Precision	Average Recall	F Score
Physics	9	0.25	0.17	0.21
Chemistry	10	0.44	0.39	0.41
Biology	15	0.66	0.46	0.54

For Text simplification we used both lexical and syntactic approaches exactly as described by De Belder and Moens (2010). The method first simplifies a given piece of text lexically. For this it uses the Wordnet [6] (Miller, 1995) to find synonyms of a term and LWLM(Deschacht et al. , 2012) to find alternate words. The intersection of the synonyms and alternate words is fed to the Kucera Francis Frequency and the the word with highest score is used in the simplified text. For syntactic simplification the appositions, relative clauses, infix and prefix subordinations and infix coordinations are removed. The FRE scores are calculated for 34 passages including the 9 passages in the user study the results are shown in 6.The FRE improves by a significant amount due to notable decrease in average sentence length.

Table 6: Results of Text Simplification



Passage Type	Average Syllables per word	Average words per Sentences	Average FRE
Original Passages	1.56	13.66	61.75
Simplified Passages	1.53	11.43	65.95

The results obtained by the methods to find out candidate words for linking are quite poor. Slight improvement is seen for biology passages but otherwise they are not so good. Other techniques like the ones proposed by Mihalcea and Csomai(2007) may be applied for automatic linking [15]. However the results obtained for Text simplification using algorithm proposed by De Belder and Moens(2010) are very good and can be used both on the links and passages, once suitable links are decided.

V. CONCLUSION AND FUTURE SCOPE

On the basis of our experiments we conclude that readability can generally be enhanced by enriching it with links to a knowledge base. If the procedure is accompanied by some lexical and syntactic text simplification the results may improve much more.

Empirical studies may be conducted to get an ideal number of links that should be provided. Other parameters that might impact readability in hyper-linked space need to be researched. If some passage is attempted a very large number of times by different users then based on the data of link usage a better and generalized readability score for that passage can be obtained. Overall the study conducted can serve as platform for further research in this domain.

REFERENCES

[1] Alinda Drury. 1972. Evaluating readability. IEEE transactions on professional communication, 28(4):11–14.

[2] BertranC.Bruce, Andee Rubin, and Kathleen S. Starr. 1981. Why readability formulas fail. IEEE transactionson professional communication,24(1):50–52.IEEE

[3] Edgar Dale, and Jeanne S.Chall. 1948. A formula for predicting readability. Educational research bulletin,11–28. JSTOR.

[4] Edward Fry. 1968. A readability formula that saves time. Journal of Reading,11(7):513–578. JSTOR.

[5] G Harry McLaughlin. 1969. SMOG grading: A new readability formula. Journal of Reading,12(8):639–646. JSTOR.

[6] George A. Miller. 1995. WordNet: a lexical database for English. Communications of the ACM,37(11):39–41. ACM.

[7] Henry Ku, Winthrop Francis Nelson, and et al. 1967. Computational Analysis of Present-Day {A}merican {E}nglish. Brown University Press.

[8] Jan De Balder, and Marie-Francine Moens. 2010. Text simplification for children. In Proceedings of the SIGIR workshop on accessible search systems,19–26.

[9] Janice C. Redish 1981. Understanding the limitations of readability formulas. IEEE transactions on professional communication,24(1):46–48. IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC 345 E 47TH ST, NEW YORK, NY 10017-2394.

[10] Jinhan Kim, Sanghoon Lee, Seung-Won Hwang, and Sunghun Kim. 2013. Enriching documents with examples: a corpus mining approach. ACM Transactions on Information Systems (TOIS),31(1):1. ACM.

[11] Koen Deschacht, Jan De Belder, and Marie-Franciene Moens. 2012. The latent words language model. Computer Speech & Language,26(5):384–409. Elsevier.

[12] Matthew Shardlow. 2014. A survey of automated text simplification. International Journal of Advanced Computer Science and Applications, 4(1).

[13] ManjiraSinha, TirthankarDasgupta, and AnupamBasu. 2014. Influence of Target Reader Background and Text Features on Text Readability in Bangla: A Computational Approach. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 345–354.

[14] Philip T. Quimlan. 1992. The Oxford psycholinguistic database. Oxford University Press,Oxford.

[15] RadaMihalcea, and Andreas Cosmoi. 2007. Wikify!: linking documents to encyclopedic knowledge. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management,233–242. ACM.

[16] R Gunning. 1952. The technique of clear writing. McGraw-Hill International, New York.

[17] RensisLikert. 1932. A technique for the measurement of attitudes. Archives of psychology.

[18] Robert West, and Jure Leskovec. 2012. Humanway finding in information networks. In Proceedings of the 21st international conference on World Wide Web, 619–628.ACM.

[19] William H. DuBay 2007. The Classic Readability Studies.Online Submission. ERIC.