Readability and Text Simplification of Educational Content as per Academic Learning Standards – The Study Plan

Chandni Goplani[1], Prof. Devang Patel[2]
[1, 2] SOCET , AHMEDABAD

Abstract- *Texts are routinely simplified for language learners with authors relying on a variety of approaches and materials to assist them in making the texts more comprehensible. Readability measures are one such tool that authors can use when evaluating text comprehensibility. Given an input text, the goal is to predict its readability level, which corresponds to the literacy level that is expected from the target reader: rudimentary, basic or advanced.*

A novel system in which the Learning content can be Managed, Organized and Delivered is the necessity for the rapidly increasing digital content. Learning standards define the specific structure of an educational program. This document reports the literature survey for the study titled 'Readability and Text Simplification of Educational Content as per Academic Learning Standards'. The document discusses the current research made in this field and the possible enhancement for the same. It highlights the motivation and future work to be conducted.

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

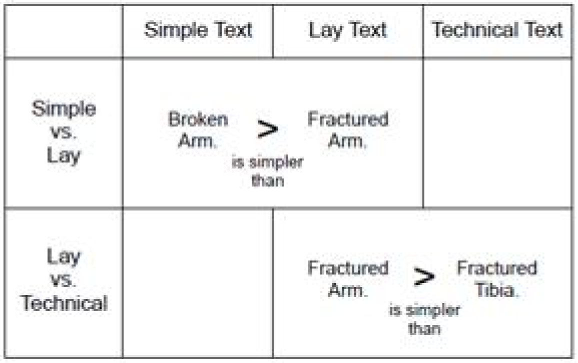
I. INTRODUCTION

"Quest for knowledge takes you to places only to find those eyes which then see what was forbidden before"

As a service provider of Information, required by an individual user it becomes pertinent that the information provided to him should also correlate to his cognitive impressions created during the part of his learning curve. The education domain is witnessing an unprecedented transformation primarily driven by digitization of vast amount of educational data and its processes. As a consequence, enormous amount of user and publisher generated learning materials are being made available online. This ever increasing amount of digitized learning material is slowly changing the way students learn, plan and progress through their educational careers. However, the widespread growth of the learning materials necessitates the development of newer systems that would efficiently manage, organize and deliver the content.[1]

Readability is the characteristic of a text that determines how easy or otherwise the text is to read and understand. Text Simplification is the process of modifying natural language to reduce its complexity and improve both readability and understandability.[2]

TABLE I. SIMPLICITY IS RELATIVE TO COMPARISON



	Simple Text	Lay Text	Technical Text
Simple vs. Lay	Broken Arm. > Fractured Arm. is simpler than		
Lay vs. Technical		Fractured Arm. > Fractured Tibia. is simpler than	

In Text Simplification, the words simple and complex are often used in relation to each other as shown in Table I

When creating simple text, we actually intend to create text which is more simple (and so less complex) than it originally was. Two other important terms to define are readability and understandability. Readability defines how easy to read a text maybe. This is typically governed by factors such as the complexity of grammar, length of sentences and familiarity with the vocabulary. Understandability is the amount of information a user may gain from a piece of text. This can be affected by factors such as the user's familiarity with the source's vocabulary, their understanding of key concepts or the time and care taken to read the text. It may be the case that a text has high readability, but low understandability. Readability and understandability are related and a text which is easier to read is likely to be more understandable, as the reader will find it easier to take the time to look over the difficult concepts. Similarly, a text which is easily understandable will encourage the reader to keep reading, even though difficult readability.

II. READABILITY FORMULAS

Readability formulas are the expressions that give a score that approximates the readability of a particular piece of text. They quantify readability. Well known readability formulas for English language are the Flesch Formulas, the Dale-Chall Formula [3], the Gunning Formula [5], the SMOG formula [6], the Fry Formula[4] and a few others. The chief parameters used by these formulas to calculate readability are average sentence length, average word length in syllables, percentage of difficult words, polysyllable count etc. Some of the well-known readability formulas are:

Flesch Reading Ease

FRE = 206.835 - 1.015* (No. of words / No. of Sentences) - 84.6 * (No. of syllables/No. of Words)

Flesch Kincaid Grade level

FKG = 0.39 * (No. of words / No. of Sentences) +11.8 * (No. of syllables/ No. of Words) - 15.59

Gunning Fog Index

GFI = 0.4 * [(No. of words / No. of Sentences)+100 * (No. of Complex Words /No. of Total Words)]
where complex words are words with three or more syllables.

Coleman Liau Index

CLI = 0.0588 * (Average number of letters per 100 words) - 0.296 * (Average number of sentences per 100 words) - 15.8

Automated Readability Index

ARI = 4.71 * (No. of Characters / No. of Words) +0.5 * (No. of Words/ No of Sentences) - 21.43

Dale-Chall Formula

Raw - Score = 0.1579 * (Percentage Of Difficult Words) +0.0496 * (Average Sentence Length) +3.6365
where difficult words are words not in Dale-Chall list of 3000 words. The Raw Score is than mapped to a predefined Grade level.

2.1 Usefulness of Readability Formulas

The readability of a text is a very important characteristic of the text especially if it is intended for a large audience. Readability formulas provide a good measure of the readability of the text and hence are frequently used by authors and publishers for evaluation and revision of text. A few uses of the readability formulas are as follows:

- Readability is very essential for school textbooks. The fact is mentioned in the guidelines given by National Council for Educational Research and Training (NCERT) for preparation of school textbooks in India. Readability formula can be used and are in fact used to classify education material based on grade level.
- They can be used by government agencies for their policy documents so that they are comprehensible by the average reader. In fact Flesch Reading Ease is used by US Department of Defense. Florida uses it for its life insurance policies. The Flesch Kincaid Grade level was specifically designed for USNavy.
- They can be used for revising medical documents and manuals for drugs.
- They can be used by search engines for retrieving documents based on reading level.

2.2 Limitations of Readability Formulas

- Though readability formulas can give essential feedback to writers and publishers, caution must be exercised while using them. Following are few aspects of writing that existing readability formulas overlook as pointed out in [14],[13]
- They cannot be used text other than prose.
- They ignore grammatical errors and syntactic simplicity
- Scores given by formulas relying on vocabulary list can be tempered with using unknown or made up words.
- They assume that smaller sentences are always more readable but sometimes they may not be as comprehensible
- User background and relevance to him are not considered.
- They work at sentence level and cannot measure how coherent the entire text is. However this limitation is addressed using Coh-Metrix [14].
- Readability formulas do not measure typographic features-illustrations, type size, typeface, use of whitespace, attractiveness of presentation-that affect how readers understand and use documents. This factors relate to legibility which is actually can be considered as a subset of readability.
- They do not consider graphs, charts, figures, tables etc. which impact readability.

III. APPROACHES

Text simplification is the process of decreasing the complexity of text both at sentence level and at word level. Text simplification generally results in improving the readability of the text and makes it better suited for a broader range of readers. Automatic text simplification takes a piece of text as an input and by some algorithm produces a simpler text. It comes under the umbrella of Natural Language Processing (NLP).The chief approaches to text simplification are lexical, syntactic and combination of both or modified hybrid approaches.[2]

3.1 Lexical Approach

In the lexical approach to text simplification difficult words are identified and are replaced by simpler and more commonly used alternatives. Generally lexical databases like WordNet [9], Kucera Francis Frequency [8],[10] etc. are used to get the more common synonyms of the complex words. WordNet keeps group of synonyms of English words known as synsets. Kucera Francis Frequency has the frequency count of about 1 million words from Brown Corpus. Larger Database of words can certainly provide better estimates of frequency. These approaches sometimes suffer from loss of meaning when the words cannot be distinguished. Hence Word Sense Disambiguation is also used. Latent Word Language model(LWLM) [11] may be used for this task. LWLM can generate semantically related words. Comparing them with alternatives given by WordNet, the word with different meaning can be removed. Finally the alternatives are ranked and the highest ranked word is used as substitute. The entire process is shown in Figure 3.1as given in [2] . Lexical Simplification can also be applied at phrase level.

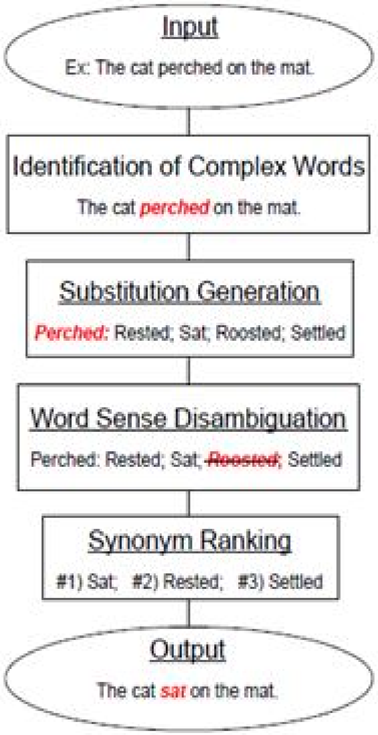

Input
Ex: The cat perched on the mat.
Identification of Complex Words
The cat perched on the mat.
Substitution Generation
Perched: Rested; Sat; Roosted; Settled
Word Sense Disambiguation
Perched: Rested; Sat; ~~Roosted~~; Settled
Synonym Ranking
#1) Sat; #2) Rested; #3) Settled
Output
The cat sat on the mat.

Figure 3.1: Lexical Simplification Process

3.2 Syntactic Simplification

In syntactic simplification the complex structure of a sentence is changed and the sentence is rewritten into one or more sentences with simpler structures. It generally involves 3 stages: analysis, transformation and generation. In analysis, a parse tree is created to realize the structure of sentence. It helps to determine if simplification is required. In transformation, based on predefined set of rules the parse tree is updated. Then further modifications are made to generate sentences having improved readability and cohesion. The process is shown in Figure 3.2 as given in [2].

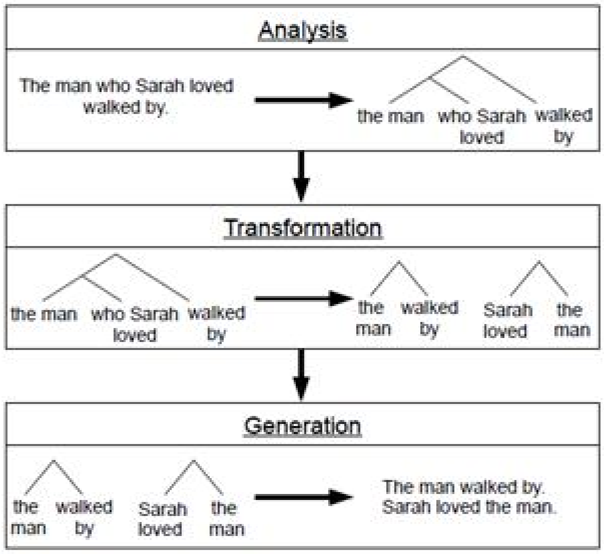


Figure 3.2: Syntactic Simplification Process

REFERENCES

[1] Danish Contractor, Kashyap Popat, ShajithIkbal, Sumit Negi, Bikram Sengupta and Mukesh Mohania. Labeling Educational Content with Academic Learning Standards. IBM Research

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

[3] W. H. DuBay. The classic readability studies. Online Submission, 2007.

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

[5] R. Gunning. The technique of clear writing, 1952.

[6] G. H. McLaughlin. Smog grading: A new readability formula. Journal of reading, 12(8):639–646, 1969.

[7] Text readability and intuitive simplification: A comparison of readability formulas by Scott A. Crossley, David B.Allen Danielle S. McNamara

[8] H. Ku,W. N. Francis, et al. Computational analysis of present-day {A}merican {E} nglish. 1967.

[9] G. A. Miller. Wordnet: a lexical database for english. Communications of theACM, 38(11):39–41, 1995.

[10] P. T. Quinlan. The Oxford psycholinguistic database. University Press, 1992.

[11] K. Deschacht, J. De Belder, and M.-F. Moens. The latent words language model. Computer Speech & Language, 26(5):384–409, 2012.

[12] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pages 233–242. ACM, 2007.

[13] B. C. Bruce, A. Rubin, and K. S. Starr. Why readability formulas fail. IEEE transactions on professional communication, 24(1):50–52, 1981.

[14] J. C. Redish. Understanding the limitations of readability formulas. IEEE transactions on professional communication, 24(1):46–48, 1981.

[15] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. Coh-metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments,& Computers, 36(2):193–202, 2004.