Book Recommendation for different age groups Using Text Readability

Jaya Ingle
cs17m060@smail.iitm.ac.in
and
Shubham Patel
cs17m051@smail.iitm.ac.in

Indian Institute of Technology, Madras

Abstract. Every text has some readability associated with it and that is dependent on certain set of features. The complex text which contain large sentence and complex grammatical structure is difficult to read. The readability of a person grow as he aged he absorbs more information and read more text. Here in this work we have taken the features mentioned by Vajjala[1], done the analysis on them, since it was not clear in the paper why they were taken and we concluded that why they worked well. Moreover we have introduce word difficulty level feature based on CEFR based dataset[6]. We have seen that they give good accuracy individually. Moreover when combined with Vajjala's features, improvement in performance is observed.

Keywords: Readablilty · Computational Lingustics · Natural Language Processing.

1 Introduction

Every text carry some form of ease with which a reader can read it. That ease is associated with number of factors. One major factor is the **age**. Different age group people have different ability to read and understand the text. A high readable text can ease the efforts of reader and his reading speed. Small age group people(i.e children) usually needs more readable text as compared to youngsters. To figure out the readability of the text, several keys play role like content of text, the way it is presented, how sentences are linked etc. Finding proper reading material according to reader's text readability is time consuming both for the teacher as well as reader. In such a case, an automated system which can select the reading material and assess the reading level of reader can help a lot.

2 Corpus

We have trained and tested our system on Weebit corpus which is formed by *WeeklyReader* magazine and *BBC-Bitesize* website. WeeklyReader magazine

contains text corresponding to age groups between 7-8, 8-9, 9-10 and 10-12 years old. The BBC-Bitesize website's data is focused for 2 grade levels 11-14 and 14 -16. Weebit is formed by merging both the data. It basically have divided the text for 5 grade scales depending upon the age. The grade levels and other details are shown in the table below.

Table 1. Details of Weebit Corpus

Grade Level	**Age group**	**Number of Articles**
Level 2	7 - 8	629
Level 3	8 - 9	789
Level 4	9 - 10	807
KS3	11 - 14	646
GCSE	14 - 16	7615

As the number of documents are not equal in each grade level, our results can be biased if we will use all of the data. So we use 500 documents for training, 100 for validation and 29 for testing in each grade level.

3 Features

We have considered 4 set of features : Traditional Features , Traditional Formula, Lexical Features, Syntactic Features. Syntactic and Lexical Feature are mainly based on Second Language Accusation study by Xiaofei Lu[4][5]. Other than this, we have added one more feature which is based on **CEFR** (Common European Framework of Reference for Languages)levels, further details are given in the subsections below.
Below are the description of each features along with the intuition behind it. Motivation behind discussing all the features here is to get an idea about their existence. Further in experiments section, it is shown that which ones are actually important.

3.1 Traditional Features

- **Number of characters per word**(NCW) [N_{char}/N_{words}] : Length of the word have major effect in deciding the reading level of the person. Kids of age group 7-8 years can find it difficult to read a long word whereas for higher age group 14-16 years it may not be difficult.

- **Number of syllables per word**(NSW) [N_{syl}/N_{words}]: Syllables signifies the pronunciation of the word. Word with large number of syllables are not easily readable by reader of smaller grade scale.
  Short words are usually preferred over long words.

Example: 'conversation' there are other words which can be used at its place like 'talk' . By using 'talk' we can lighten his reading load.

3.2 Traditional Formula

There were two traditional Formula suggested by Sowmya Vajjala[1]

- **Flesch-Kincaid Grade level score**

$$FleschKincaid = 0.39 \times \frac{total\ words}{total\ Sentence} + 11.8 \times \frac{total\ Syllable}{total\ word} - 15.59$$

This readability score has been used extensively. Even Microsoft office provide this score as a statistics to judge readability. It consider the combination of lexical features while ignoring syntactical features.

- **Coleman-Liau readability formula**

$$CLI = 0.00588 * L - 0.296 * S - 15.8$$

- L is the average number of letters per 100 words.
- S is the average number of sentences per 100 words.

The more the number of characters in a word, the longer the words. And generally the longer words are more complex in term of their sense. They often represent a collective information of much more simpler word. So, if the no. of characters per words is increasing for text, it can be inferenced that the complexity of the sentence in term of readability is going to rise.
The larger the magnitude of average no. of sentence per 100 words, the simpler will be the text. If a series of 100 words contain more sentence that indicate that no. of sentences are increased and sentences are of smaller length. Smaller length sentence are largely simpler than the large length sentences. Simpler sentences increase readability for the lower age group people.

- **Other Readability formulas** We consider few more readability formulae like
  - Automated Readability Index

$$ARI = 4.71 \times \frac{Characters}{Words} + 0.5 \times \frac{Words}{Sentences} - 21.43$$

  - Dale Chall readability score The Dale-Chall Formula is different from other because other formulas use word-length to compute text difficulty while the Dale-Chall Formula focus on count of hard words. The Dale-Chall Formula calculates the US grade level of a text sample based on sentence length and the number of hard words. These hard words are words that do not appear on a specially designed list of common words familiar to most 4th-grade students.

$$RawScore = 0.1579 * (PDW) + 0.0496 * ASL$$

$$PDW = Percentage\ of\ Difficult\ Words$$

$$ASL = Average\ Sentence\ Length\ in\ words$$

If (PDW) is greater than 5%, then:
Adjusted Score = Raw Score + 3.6365,
otherwise
Adjusted Score = Raw Score

So, it can be concluded that the above scores should do fine. Later we see in the experiment sections how they actually performed compare to our intuition given Weebit corpus.

3.3 Lexical Features

- **Type-token ratio(TTR) [T/N]**: It is the ratio of vocabulary size(T) to the length of the text(N). It helps in finding how much diverse is the vocabulary.
  The larger the resulting TTR the less repetitive the vocabulary usage
  It have been observed in several papers that TTR is highly text length dependent: longer the text, less chances of introduction of new words which results in lower TTR. So several modifications has been tried to solve text size problems, listed below:

  - Root type token ratio (RTTR) [T/sqrt(N)]
  - Corrected Type token ratio (CTTR) [T/sqrt(2N] - It proves to be most predictive feature.
  - Bi-logarithmic type token ratio (BTTR) [log(TTR)]
  - Uber Index ($log(T)^2/log(N/T)$)

- **Lexical Density** (LD) N_{lex}/N: Lexical density is defined as the number of lexical words N_{lex} (nouns, adjectives, verbs, adverbs) divided by the total number of tokens(N) in the text.
  It is actually telling how many words out of total words are giving text its actual meaning.

- **Lexical Variations**: The ratio of the number of lexical types to lexical tokens.
  - *Noun Variations*(NV) [N_{noun}/N] : Proportion of nouns divided by number of tokens.
    We are always interested in 3-dimensional world. If you give a reader something that he can see, visualize and touch, it wakes him up.
    Names(people) matters like Harry potter, Disney, etc.Nouns are the drivers of the sentence. They can maker reader more vivid.

- *Adjective Variations*(AdjV)[N_{adj}/N]: Proportion of adjectives divided by number of tokens.
  Adjective gives description of noun.They modify and add feelings/qualities to the noun. But at the same time, they are difficult to interpret for kids, as they don't know meanings of such words.

- *Adverb Variations*(AdvV) [N_{adv}/N]: Proportion of adverbs divided by number of tokens.
  Many researchers believe that adverbs should be use as less as possible.They reduce the reading speed.
- *Verb Variations*(VV)[N_{verb}/N] : Verbs are usually used to describe an action, state, or occurrence. Verbs are the engine of the sentence Basically verbs(active) requires actor to do the action. So reader can see the things happening in their heads. Several variations of verb variations such Squared verb variations(svv) and corrected verb variation(cvv) has been also been tried

- *Academic Word List(AWL)* : The proportion of words that are the part of academic word list.
  The AWL is a list of words which appears with higher frequency in the English-language academic text.List consist of 570 word families and it is divided into 10 sublists. Sublist 1 consists of the 60 most common words in the AWL. It is created bu Averil Coxhead, Victoria University of Wellington, New Zealand.[7]

3.4 Syntactic Features

– **Mean Length of**
  - **Sentence(MLS)**
    Sentence is a basic unit of English Grammar. A sentence convey a complete thought.
    Intuition Length of a sentence can be seen as correlating with age. Longer sentences often have longer context to retain in memory, which is difficult from a child's perceptive. But a adult who have adequate experience in reading can be comfortable with reading long sentences. So, average length of a sentence can be a good feature.

  - **Clause (MLC)** Sentence can be further broken down into clauses. A sentence can either consist of a single or multiple clauses. A clause is a grammatical structure which has a verb and a subject.[10]
    Intuition The sentence with multiple clauses are more complex as compare to the sentence with the single clause. Sentence with simple clauses have less information to keep in the back of the mind while reading, which indicate that it will be more feasible for less experience reader. But there's catch here. We are taking feature as mean length of clause. And a sentence can often divided into multiple small clauses which tends

to give contradicting estimate corresponding to above analysis. So, it can be thought as that mean length of clause should not be a effective parameter. Since it can have multiple meaning which can also be misleading.

- **T-unit(MLT)** A T-unit is defined as the main clause with all the subordinate clauses and the construction that go with it. The mean length of T unit is the sum of length of all T units divided by the total T units. T unit is often define in the linguistic community as the fundamental yardstick for measuring syntactic development.[9]
  Intuation Word/Sentence often fails to capture the complexity of a sentence in term of age. Since a very young child can write a very long sentence consist of several compound main clause. So, a t-unit which focus on the main clauses instead of a sentence or the individual clauses seems to provide better picture for the age group.

– **Number of Clauses/T-Units** The feature "no. of clauses per T units" can be seen like this as a T unit is consist of several clauses. So, this give us the average estimate of clauses per unit. As the no. of clauses increases, it is a sign of increase in the complexity of the text. Also such text can be longer. So, the higher this quantity is the higher age group should be assigned to it.

– **Num. of Complex-T-Units per T-unit (CT/T)** Complex T unit are defined as any T unit that contain the dependent clause. So it is the parameter that suggest on an average how many t-units contain the dependent clause. The dependent clause further increase the length of the sentence. A dependent clause can often contain some important information which also be taken into the consideration. So, it can be concluded that the average age of the reader can become higher because of these factor. A experience reader will do well with such sentence as compare to the one that doesn't have experience.

– **Dependent Clause to Clause Ratio (DC/C)** The measure is similar to the above measure but now we are focusing on the clause instead of the T units. Let's analyze it. Though no. of clauses alone does not provide any significant information. But with dependent clauses, it is capturing in a way same information as CT/T with different normalization.

– **Dependent Clause to T-unit Ratio (DC/T)** We have a similar feature above CT/T. But a complex T unit can have multiple dependent clause. So, CT/T can be consider as the lower bound for the DC/T. Here we are trying the capture the similar ratio as the one that we have done in the DC/C, but with the more fundamental unit.

– **Co-ordinate Phrases per Clause (CP/C)** Co-ordinate are words in English grammar that joins two element of the language. For e.g. and, but. So, Co-ordinate Phrases per Clause give an idea about no of co-ordinators that

are used on an average. The increase in use of CP per C also indicates that sentence is complex.

– **Co-ordinate Phrases per T-unit (CP/T)** The similar kind of argument can be made with respect to co-ordinate per T-Units. We are just changing the normalization criteria here. It can be possible some extra information that is missing in the previous features.

– **Complex Nominals per Clause (CN/C), Complex Nominals per T-unit (CN/T)** The nominal in the English language are the group of words which describe or represent an Entity.[11] Complex nominals are consist of [12]
  1. nouns plus adjective, possessive, prepositional phrase, relative clause, participle or appositive.
  2. nominal clauses
  3. gerunds and infinitives in subject positions.

  The complex Nominals per Clause and Complex Nominals per T units are again can also be helpful in measuring the complexity of the sentence.

– **Verb phrases per T-unit (VP/T)** Verb Phrase is a part of a sentence that contains the verb and any direct or indirect objects but not the subject. Verb phrase e.g. She was walking quickly to the mall. where was walking is a verb phrase. So, Avg no of verb phrase in a T unit also lead to the way of complexity of the sentence. The T-unit with multiple verb phrase will be complicated as compare to the with lower.

– **Other syntactic Features**
  - Number of Noun/Verb/Preposition Phrase per sentences. The idea is remain the same. The more the Noun/Verb/Preposition the more complex and less readable sentence will be.
  - Average length of NP,VP,PP. If avg length of Noun Phrase, Verb Phrase, and Preposition Phrase will be more sentence tends to be more complex and hard to read.
  - Number of Dependent Clauses/per sentence, Complex-T units/per sentence, Co-ordinate Phrases/per sentence, SBARs/per sentence. Intuition largely remains the same.
  - Average height of Parse tree. The higher the average height of parse tree, the complex will be the sentence.

3.5 Other Feature

Word Difficulty Level/S It measure based on the
https://goo.gl/zBszfm. So there are 6 level in CEFR exam. In the dataset mentioned above teachers are asked to classify English words according to the level

of the exam. A mean score is calculated for each word collectively given by all the teachers. We have taken the ceil of such score and classify the words again into six distinguish classes. And we have calculated the score of no. of words that belongs to a particular level divide by total sentence. We hope that it will capture the difficulty score nicely.

The idea here is that we are not sure what particular variations of features actually work well. We just have a intuition that these set of feature should work better. And it can be possible that some particular information is missed by the one variation of a feature that might be captured by the another. There is one more intuition behind going for so many variations is that many of the features that we have taken cannot be learned accurately and precisely. So, taking multiple variations of multiple features which form foundation of syntax, kind of provide a safety net for the falls of missing essential information. The soul purpose of all the features is capture the complex nature of text with help of lexical, linguistic and difficulty features associated with the text. In a glance so many features seems to be disturbing. But if see it as a perceptive in how we read, understand and learn, we can find that we only see a part of the structure of the text and a larger part that is hidden is captured by our brain. And we often term that hidden representation which we have learned as intuition. So, there are some obvious feature that we can see, but there are so many more which we actually absorb.

4 Methodology

4.1 Feature Extraction

We have use Stanford parser and tagger for the majority of the feature extraction. Also, we have use NLTK(python) library for interfacing with Stanford parser. We have used NLTK sentence tokenizer to tokenize text into sentences and word tokenizer to tokenize text into sentences. We have also used Lu. tools to extract lexical syntactical features[8], which provide all the syntactic features mentioned in SLA literature. Moreover we have use python textstat library to extract classical formula and some new formulas that we have introduced in this report.

4.2 Models

The goal here is to classify a text according to it's readability among 5 classes. For this purpose we have considered 3 key methods that are Support Vector Machine(SVM), Logistic Regression and Multilayer Perceptron Classifier(MLP). We have used the python sklearn library which consists of all of these methods. We have trained the models on the 2500 entries, 500 from the each class. We have created a validation set which consist of 29 entries from each class, so that we can not be perform biased towards our goal of achieving good result on unknown data. Then in the end we have calculated test accuracy. Which we have reported in the results.

5 Experiments

For experiments we try different set of features using different combination of traditional(also formulae), lexical and syntactic features.As said above, we use model SVM, MLP and logistics regression. Performance of MLP was comparatively good than SVM and logistic regression, that's why we only show experiments results of MLP. MLP was hard to tune as it has so many features, comparatively was easy to tune. One intuition behind it that MLP is a true classifier as compare to other two models. The SVM and logistic regression on the other hand use one vs all or one vs rest method. And also MLP are more capable to learn more complex function.

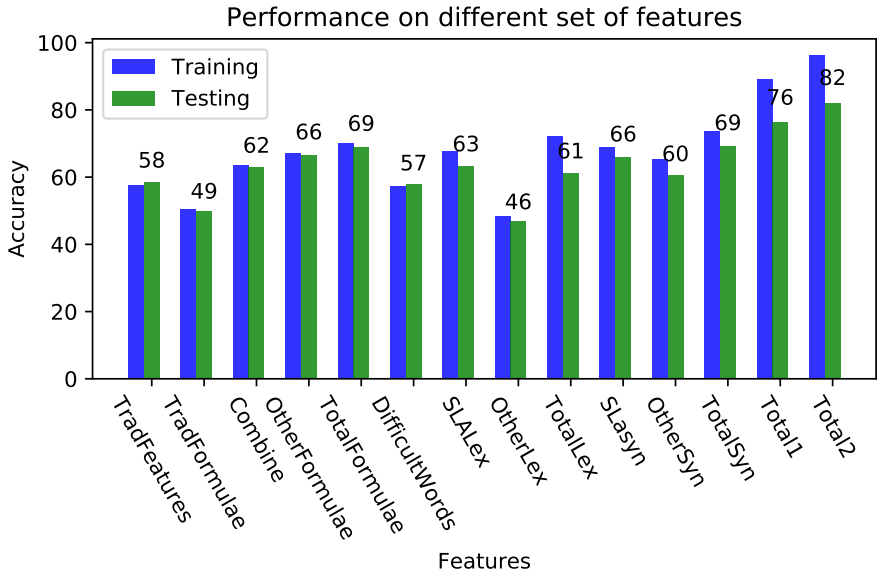

Fig. 1. Using MLP

In the bar plot above, we use certain shortform for features. The table below describes what are those features and the number of features in that particular set of feature.

Table 2. Explanation of set of features

xlabels	**Explanation**	**Number**
TradFeatures	Traditional Features- NSW, NLW and no. of words per sentence(same as MLS)	3
TradFormulae	Traditional Formulae such as Coleman li's and Flesh Kincard score	2
Combine	Combination of above 2 set of features	5
Otherformulae	Some other formulae as described in other formula section	8
TotalFormulae	Total= TraditionalFormulae + OtherFormule	8+2 =10
DifficultWords	Difficult word list from CEFR exam	6
SLALex	Lexical Features from SLA research	13
OtherLex	Other Lexical Features includes NCW, Academic WordList,NSW	3
TotalLex	Total is the combination of SLAlexical +OtherLexical Fetures	13+3 = 16
SLAsyn	Syntactic features from SLA research	14
OtherSyn	As described in other syntactic Features section	11
TotalSyn	Combination of SLAsyntactic Features and OtherSyn	14+11 =25
Total1	Total features taken in Sowmya and Vajjala's paper	25+16 =41
Total2	Total which on adding OtherFormulae+Difficult word	41+10+6

5.1 Why taking all variations of few features was fruitful?

Table 3. Performance of different variation of ttr

Features	**Training Accuracy**	**Testing Accuracy**
ttr	42.2	42.4
bttr	33.5	32.8
cttr	28.96	31
rttr	33.5	32.8
uber	28.16	27
{ttr,rttr}	42.5	43
{bttr,cttr}	43.2	42.6
{uber,ttr}	41.8	43
{ttr,cttr,rttr,bttr,uber}	**44.16**	**44.6**

Table 4. Performance of different verb variation

Features	**Training Accuracy**	**Testing Accuracy**
VV1	28.24	28
svv	27.16	27.6
cvv	29.04	27.8
{VV1,svv}	38.24	36.8
{svv,cvv}	30	27
{VV1,svv,cvv}	**39**	**38**

Table 5. Performance of Dependent Clause per Sentence/clause/t-unit

Features	**Training Accuracy**	**Testing Accuracy**
DC/T	36.3	37.4
DC/C	37.24	39.6
DC/S	36.3	39
{DC/T,DC/S}	39.24	39.6
{DC/C,DC/S}	43.25	43.6
{DC/C,DC/T}	44.6	46
{DC/C,DC/S,DC/T}	**48.4**	**46.6**

Initially it may looks like taking all variations of same features may degrade the performance. But on experiments, we get some different results, rather than dropping, it increases the performance. So the reason we discussed above proves true that taking all the features proves a good idea.

5.2 Observations and Inference

- **All the set of features performs good**
  If we see and compare to the baseline accuracy that is 20% (100/5)here. All of the set of features have performed relatively well. This gives a kind of validation that the features selected are effective. And that's a good sign. This is also an indication that simple features such as traditional features actually performs well.
- **Traditional Features performs bad as compared to lexical and syntactic features**
  Traditional features itself will not be able to capture the structure and meaning of the sentence. It also includes number of syllables per word which can not be precisely determine in all cases.
- **Traditional Formulae when combined other formulae performs pretty good, but it not much improvement over other formulae.**
  It is because some other formula also count number of complex words rather simply counting number of words per sentence and number of characters in word
- **Different variation of ttr when used collectively performed better**
  Since ttr is complex, and a single variation is not enough to capture the effect of ttr. Thus, multiple variable performs effectively.
- **Difficult word feature alone was able to get 57% accuracy**
  The 6 levels of difficult words that we have taken indeed perform well. This result is as we are expecting. This is also can be seen as the more the abstraction a word provide the more it increase the readability difficulty.
- **SLASyntactic feature when combined with other syntactic features increase the overall accuracy**
  Since other syntactic features capture much more variation of structural information which can further provide new information. So, result in overall better result.

– **When taken all the features together we get the best result**
So, when we have taken all the features we get the best result. That shows that our all the features has contributed positively to our goal of computing text readability.

6 Results and Conclusions

We have get best test results when we have taken all of our proposed features together, which is a indication of validation of the features that we have selected. We got around 82% (Original Paper has 93%). But there's a catch we don't have the exact corpus that Vajjala[1] has used. By taking the same set of features that original papers author used we got around 76% accuracy. Our 82% accuracy is 6% improvement over that. Which indicate that new features Difficultly level of Words, Dale-Chall Score helped us to improve our performance.

7 Future Work

The above experiment is performed in the context of Weebit which is consist of magazines which is published for native English speaker. The performance of the feature may vary for the non-native English speaker. To show that above features also perform well for non-native English text, the cross classification of NCERT books can be done for that. Moreover a new completely new model can be train for the NCERT books and given that we can also see that how well that adopt on native-english corpora such as Weebit.

For the classification of children books solely, a corpora can be build consist of children books that are arrange in age wise category. These features can further be used to train a model given that corpora. This can be a effective way to suggest books to a kid belong to certain age.

The text readability can also be seen in a different aspect other than linguistic features. We require much more than the linguistic features to capture the readability aspect of text. As a human whenever we read some thing, we relate it to hidden representation that we have for that text. Which we have learned through events of our life. That representation most of the time plays a crucial role in our understanding. But the main difficulty with that aspect is that it can vary from person to person. But a rough estimate can be made considering each student that study in a particular grade may have a similar representations. Those representation can be further learned by using the compulsory text that a student has go through to reach that grade. A neural Network can play a significant role in learning that representation. What can be done further from here can part of some other study, for some other time.

References

1. Sowmya Vajjala, Detmar Meurers. *On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition.* The 7th Workshop on the

Innovative Use of NLP for Building Educational Applications, pages 163173,Montreal,Canada, June 3-8, 2012/
2. Sowmya Vajjala, Detmar Meurers. *On The Applicability of Readability Models to Web Texts*. Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, pages 5968, Sofia, Bulgaria, August 4-9 2013
3. Brian Richards. *Type/Token Ratios: what do they really tell us?*
4. Lu, Xiaofei (2010). `Automatic analysis of syntactic complexity in second language writing`. International Journal of Corpus Linguistics, 15(4), 474-496.
5. Lu, Xiaofei (2012). `The relationship of lexical richness to the quality of ESL learners' oral narratives`
6. Guzey, Onur, Sohsah, Gihad, and Unal, Muhammed, Classification Of Word Levels With Usage Frequency, Expert Opinions And Machine Learning, 2014.
7. Vocabulary Exercises for the Academic Word List
`http://www.englishvocabularyexercises.com/AWL/`
8. L2 Syntactic Complexity Analyzer, Xiaofei Lu
`http://www.personal.psu.edu/xxl13/downloads/l2sca.html`
9. Definitions of the "T-unit"
`http://www.kissgrammar.org/Essay/Essay009_Def_TUnit.htm`
10. Clauses : What are clauses ?
http://www.grammar-monster.com/glossary/clause.htm
11. Nominal Group (Functional Grammar)
https://en.wikipedia.org/wiki/Nominal_group_(functional_grammar)
12. Complex nominals
`https://goo.gl/by4GJT`
13. SKLearn : SVM Tool
`http://scikit-learn.org/stable/modules/svm.html`
14. Using Support Vector Machines Effectively
`https://neerajkumar.org/writings/svm/`
15. SKLearn : Preprocessing of Data
`https://goo.gl/Kq8sMM`
16. LNCS Homepage, http://www.springer.com/lncs. Last accessed 4 Oct 2017
17. Gunning fog index
`https://en.wikipedia.org/wiki/Gunning_fog_index/`
18. Automatic Readability Index
`https://en.wikipedia.org/wiki/Automated_readability_index`