Automated Reading Comprehension Clustering

Charles Hathaway, David Hedin

December 14, 2015

Abstract

Determining the level of readability of documents, especially books, has lots of application in the domain of education. It helps to quantify and group books which may be used at a particular reading level, therefore enhancing the classroom expierence for both instructors and teachers. In this paper, we ran an experiment with a variety of books obtained from Project Gutenberg [1] organized by the project maintainers into a 2 groups; books for children, and adult fiction. To further enhance the analysis of this project, we also used our features to try and cluster books into clusters defined by the FleschKincaid readability metrics [2].

1 Introduction

There are several systems currently used to classify books based on reading comprehension level, for numerous applications ranging from selecting books for classrooms, to measuring an individuals literacy skills for both medical (autism, dyslexia, etc.) and educational purposes. In this paper, we analyze the results of utilizing several existing features to classify works in addition to a variety of novel features we created. The primary goal is to cluster books in groups representing the original designation of books; adult fiction and children fiction.

Given an input of 2292 books (320 children, 2002 adult, with 32 books overlapping) we achieved an F-score of 87.5%; this is a significant improvement over the 62% baseline [1]

2 Previous Works

Although not extensively, we did evaluate and learn from a number of previous works. Most significantly, we borrowed features from work done by Feng at al. [3].

[1]There was some in-class discussion which suggested our baseline would be (total number of adult books)/(total number of books), which would put the baseline at around 86%. After testing this experimentally, and reasoning things out, we concluded the true baseline would be 62% as our algorithm had no idea what the sizes of the clusters were, and a truly random distribution would but half in each cluster, with one cluster having a higher chance of being correct than the other

3 Methodology and system design

4 Linguistic Features

Average number of words per sentence
Average number of syllables per word
Percentage of words with more than 3 syllables
Average number of noun phrases per sentence
Average number of common and proper nouns per sentence
Average number of verb phrases per sentence
Average number of adjectives per sentence
Average number of conjunctions per sentence
Average number of prepositional phrases per sentence
Total number of noun phrases in document
Total number of common and proper nouns in document
Total number of verb phrases in document
Total number of adjectives in document
Total number of conjunctions in document
Total number of prepositional phrases in document
Number of entity mentions in document
Number of unique entities in document
Average number of entity mentions per sentence
Average number of unique entities per sentence

Table 1: List of possible features from previous work

Average word length in document
Total number of unique words in document
Ratio of unique words to total number of words in document
Ratio of proper nouns to common nouns in document
Length of document
Average number of proper nouns per sentence
Total number of proper nouns in document
Total number of passive sentences in document
Average number of prepositional phrases per sentence
Total number of prepositional phrases in document

Table 2: List of possible new features

5 Results

6 Conclusion

References

[1] (Dec. 12, 2015). Project gutenberg, [Online]. Available: `http://www.gutenberg.org/` (visited on 12/12/2015).

[2] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," DTIC Document, Tech. Rep., 1975.

[3] L. Feng, N. Elhadad, and M. Huenerfauth, "Cognitively motivated features for readability assessment," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 2009, pp. 229–237.