Ranking-based readability assessment for early primary children's literature

Yi Ma, Eric Fosler-Lussier
Dept. of Computer Science & Engineering
The Ohio State University
Columbus, OH 43210, USA
may,fosler@cse.ohio-state.edu

Robert Lofthus
Xerox Corporation
Rochester, NY 14604, USA
Robert.Lofthus@xerox.com

Abstract

Determining the reading level of children's literature is an important task for providing educators and parents with an appropriate reading trajectory through a curriculum. Automating this process has been a challenge addressed before in the computational linguistics literature, with most studies attempting to predict the particular grade level of a text. However, guided reading levels developed by educators operate at a more fine-grained level, with multiple levels corresponding to each grade. We find that ranking performs much better than classification at the fine-grained leveling task, and that features derived from the visual layout of a book are just as predictive as standard text features of level; including both sets of features, we find that we can predict the reading level up to 83% of the time on a small corpus of children's books.

1 Introduction

Determining the reading level of a text has received significant attention in the literature, dating back to simple arithmetic metrics to assess the reading level based on syllable counts (Flesch, 1948). In the computational linguistics community, several projects have attempted to determine the grade level of a text (2nd/3rd/4th/etc). However, the education community typically makes finer distinctions in reading levels, with each grade being covered by multiple levels. Moreover, there are multiple scales within the educational community; for example 1st grade is approximately covered by levels 3–14 on the Reading Recovery scale,[1] or levels C to H in the Fountas and Pinnell leveling system.[2]

For grade-level assessment, classification and regression approaches have been very promising. However, it is not clear that an increased number of classes will allow classification techniques to succeed with a more fine-grained leveling system. Similarly, regression techniques may have problems if the reading levels are not linearly distributed. In this work, we investigate a ranking approach to book leveling, and apply this to a fine-grained leveling problem for Kindergarten through 2nd grade books. The ranking approach also allows us to be more agnostic to the particular leveling system: for the vast majority of pairs of books, different systems will rank the levels of the books the same way, even if the exact differences in levels are not the same. Since most previous work uses classification techniques, we compare against an SVM multi-class classifier as well as an SVM regression approach.

What has not received much attention in recent research is the visual layout of the page. Yet, if one walks into a bookstore and rummages through the children's section, it is very easy to tell the reading level of a book just by thumbing through the pages. Visual clues such as the number of text lines per page, or the area of text boxes relative to the illustrations, or the font size, give instant information to the reader about the reading level of the book. What is not clear is if this information is sensitive enough to deliver a fine-grained assessment of the book. While

[1] http://www.readingrecovery.org
[2] http://www.fountasandpinnellleveledbooks.com

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 548–552,
Montréal, Canada, June 3-8, 2012.

publishers may have standard guidelines for content providers on visual layout, these guidelines likely differ from publisher to publisher and are not available for the general public. Moreover, in the digital age teachers are also content providers who do not have access to these guidelines, so our proposed ranking system would be very helpful as they create reading materials such as worksheets, web pages, etc.

2 Related Work

Due to the limitations of traditional approaches, more advanced methods which use statistical language processing techniques have been introduced by recent work in this area (Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005; Feng et al., 2010). Collins-Thompson and Callan (2004) used a smoothed unigram language model to predict the grade reading levels of web page documents and short passages. Heilman *et al.* (2007) combined a language modeling approach with grammar-based features to improve readability assessment for first and second language texts. Schwarm/Petersen and Ostendorf (2005; 2009) used a support vector machine to combine surface features with language models and parsed features. The datasets used in these previous related works mostly consist of web page documents and short passages, or articles from educational newspapers. Since the datasets used are text-intensive, many efforts have been made to investigate text properties at a higher linguistic level, such as discourse analysis, language modeling, part-of-speech and parsed-based features. However, to the best of our knowledge, no prior work attempts to rank scanned children's books (in fine-grained reading levels) directly by analyzing the visual layout of the page.

3 Ranking Book Leveling Algorithm

Our proposed method can be regarded as a modified version of a standard ranking algorithm, where we develop a leveling classification by first ranking books, and then assigning the level based on the ranking output. Given a set of leveled books, the process to generate a prediction for a new target book involves the following two steps.

In the first step, we extract features from each book, and train an off-the-shelf ranking model to minimize the pairwise error of books. During the test phase (second step), we rank all of the leveled training books as well as the new target (test) book using the trained ranking model. The predicted reading level of the target book then can be inferred from the reading levels of neighboring leveled books in the rank-ordered list of books (in our experiment, we take into account a window of three books above and below the target book with reading levels weighted by distance). Intuitively, we can imagine a bookshelf in which books are sorted by their reading levels. The ranker's prediction of the reading level of a target book corresponds to inserting the target book into the sorted bookshelf.

4 Data Preparation

4.1 Book Selection, Scanning and Markup

We have processed 36 children's books which range from reading level A to L (3 books each level). The golden standard key reading levels of those books are obtained from Fountas and Pinnell leveled book list (Fountas and Pinnell, 1996) in which letter A indicates the easiest books to read and letter L identifies more challenging books; this range covers roughly Kindergarten through Second Grade. The set of children's books covers a large variety of genres, series and publishers.

After seeking permission from the publishers,[3] all of the books are scanned and OCRed (Optical Character Recognized) to create PDF versions of the book. In order to facilitate the feature extraction process, we manually annotate each book using Adobe Acrobat markup drawing tools before converting them into corresponding XML files. The annotation process consists of two straightforward steps: first, draw surrounding rectangles around the location of text content; second, find where the primary illustration images are and mark them using rectangle markups. Then the corresponding XML can be generated directly from Adobe Acrobat with one click on a customized menu item, which is implemented by using Adobe Acrobat JavaScript API.

[3] This is perhaps the most time-consuming part of the process.

# of partitions	1	2	3	4
±1 *Accuracy %*				
SVM Ranker	72.2	69.4	80.6	**83.3**
SVM Classifier	47.2	61.1	55.6	**63.9**
SVM Regression	72.2	61.1	58.3	58.3
Flesch-Kincaid	30.6	30.6	30.6	19.4
Spache	27.8	13.9	13.9	11.1
Average leveling error ± *standard deviation*				
SVM Ranker	1.00 ± 0.99	1.03 ± 0.91	0.94 ± 0.83	0.92 ± 0.73
SVM Classifier	2.00 ± 1.60	1.86 ± 1.69	1.78 ± 1.57	1.44 ± 1.23
SVM Regression	1.14 ± 1.13	1.25 ± 1.11	1.33 ± 1.22	1.36 ± 1.22
Flesch-Kincaid	3.03 ± 2.21	3.03 ± 2.29	3.08 ± 2.31	3.31 ± 2.28
Spache	4.06 ± 3.33	4.72 ± 3.27	4.83 ± 3.34	5.19 ± 3.21

Table 1: Per-book (averaged) results for ranking versus classification, reporting accuracy within one level and average error for different numbers of partitions

4.2 Feature Design

4.2.1 Surface-level Features

We extract a number of purely text-based features that have typically been used in the education literature (e.g., (Flesch, 1948)), including:

1. Number of words; 2. Number of letters per word; 3. Number of sentences; 4. Average sentence length; 5. Type-token ratio of the text content.

4.2.2 Visually-oriented Features

In this feature set, we include a number of features that would not be available without looking at the physical layout of the page; with the annotated PDF versions of the book we are able to extract:

1. Page count; 2. Number of words per page; 3. Number of sentences per page; 4. Number of text lines per page; 5. Number of words per text line; 6. Number of words per annotated text rectangle; 7. Number of text lines per annotated text rectangle; 8. Average ratio of annotated text rectangle area to page area; 9. Average ratio of annotated image rectangle area to page area; 10. Average ratio of annotated text rectangle area to annotated image rectangle area; 11. Average font size.

The OCR process provides some of this information automatically; while we have manually annotated rectangles for this study one could theoretically use the OCR information and vision processing techniques to extract rectangles automatically.

5 Experiments

5.1 Ranking vs. Classification/Regression

In this experiment, we look at whether treating book leveling as a ranking problem is promising compared to using classification/regression techniques. Besides taking a whole book as input, we also experiment with partitioning each book uniformly into 2, 3, or 4 parts, treating each sub-book as an independent entity. We use a leave-n-out paradigm – during each iteration of the training (iterated through all books), the system leaves out all n partitions corresponding to one book and then tests on all partitions corresponding to the held-out book. By averaging the results for the partitions of the held-out book, we can obtain its predicted reading level.

For ranking, we use the $\mathrm{SVM}^{\mathrm{rank}}$ ranker (Joachims, 2006), which learns a (sparse) weight vector that minimizes the number of swapped pairs in the training set. The test book is inserted into the ordering of the training books by the ranking algorithm, and the level is assigned by averaging the levels of the books above and below the order. To compare the performance of our method with classifiers, we use both $\mathrm{SVM}^{\mathrm{multiclass}}$ classifier (Tsochantaridis et al., 2004) and $\mathrm{SVM}^{\mathrm{light}}$ (with regression learning option) (Joachims, 1999) to determine the level of the book directly. All systems are given the same set of surface text-based and visual-based features (Sections 4.2.1 and 4.2.2) as input.

# of partitions	1	2	3	4
± 1 *Accuracy %*				
All Features	72.2	69.4	80.6	83.3
Surface Features	61.1	**63.9**	58.3	61.1
Visual Features	72.2	72.2	72.2	**83.3**
Average leveling error $\pm$ *standard deviation*				
All Features	1.00 ± 0.99	1.03 ± 0.91	0.94 ± 0.83	0.92 ± 0.73
Surface Features	1.42 ± 1.18	1.28 ± 1.00	1.44 ± 0.91	1.28 ± 1.11
Visual Features	1.03 ± 0.88	0.94 ± 0.86	1.03 ± 0.81	0.89 ± 0.82

Table 2: Per-book (averaged) results for all, surface-only, and visual-only features, reporting accuracy within one level and average error for different numbers of partitions

We score the systems in two ways: first, we compute the accuracy of the system by claiming it is correct if the book level is within ± 1 of the true level.[4] The second scoring method is the absolute error of number of levels away from the true value, averaged over all of the books.

As we can observe from Table 1, our ranking system constantly beats the other two approaches (the ranker is statistically significantly better than the classifier at $p < 0.05$ level – figures in bold). One bit of interesting discovery is that SVM regression needs more data in order to have reliable results, as the performance is downgraded when the number of partitions goes up; the ranking approach benefits from averaging the increasing number of partitions.[5]

All three methods have the same style of learner (support vector learning), which suggests that the performance gain is due to using a ranking criterion in our method. Therefore we believe ranking is likely a more effective and accurate method than classification for this task.

One might also wonder how a traditional measure of reading level (in this case, the Flesch-Kincaid (Flesch, 1948) and Spache (Spache, 1953) Grade Level) would hold up for this data. Flesch-Kincaid and Spache predictions are linearly converted from calculated grade levels to Fountas-Pinnell levels; all of the systems utilizing our full feature set outperform these two baselines by a significant amount on both ± 1 accuracy and average leveling error.

5.2 Visual vs. Surface Features

In order to evaluate the benefits of using visual cues to assess reading levels, we repeat the experiments using $\mathrm{SVM}^{\mathrm{rank}}$ based on our proposed ranking book leveling algorithm with only the visual features or only surface features.

Table 2 shows that the visual features surprisingly outperform the surface features (statistically significant at $p < 0.05$ level – figures in bold) and on some partition levels, visual cues even beat the combination of all features. We note, however, that for early children's books, pictures and textual layout dominate the book content over text. Visual features can be as useful as traditional surface text-based features, but as one moves out of primary literature, we suspect text features will likely be more effective for leveling as content becomes more complex.

6 Conclusions

In this paper, we proposed a ranking-based book leveling algorithm to assess reading level for children's literature. Our experimental results showed that the ranking-based approach performs significantly better than classification approaches as used in current literature. The increased number of classes deteriorates the performance of classifiers in a fine-grained leveling system. We also introduced visual features into readability assessment and have seen considerable benefits of using visual cues. Since our target data are children's books that contain many illustrations and pictures, it is quite reasonable to utilize visual content to help predict a more accurate reading level. Future studies in early childhood readability need to take visual content into account.

[4] Note that this is still rather fine-grained as there are multiple book levels per grade level.

[5] We only partition the books up to 4 sub-books because the shortest book we have only contains 4 PDF pages (8 "book" pages) and further partitioning the book will lead to sparse data.

References

K. Collins-Thompson and J. Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT / NAACL 2004*, volume 4, pages 193–200, Boston, USA.

L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 276–284, Beijing, China. Association for Computational Linguistics.

R. Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221–233.

I. Fountas and G. Pinnell. 1996. *Guided Reading: Good First Teaching for All Children*. Heinemann, Portsmouth, NH.

M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL HLT*, pages 460–467.

T. Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA.

T. Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226. ACM.

S. Petersen and M. Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106.

S. Schwarm and M. Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics.

G. Spache. 1953. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413.

I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, page 104. ACM.