On The Applicability of Readability Models to Web Texts

Sowmya Vajjala **Detmar Meurers**
Seminar für Sprachwissenschaft
Universität Tübingen
{sowmya,dm}@sfs.uni-tuebingen.de

Abstract

An increasing range of features is being used for automatic readability classification. The impact of the features typically is evaluated using reference corpora containing graded reading material. But how do the readability models and the features they are based on perform on real-world web texts? In this paper, we want to take a step towards understanding this aspect on the basis of a broad range of lexical and syntactic features and several web datasets we collected.

Applying our models to web search results, we find that the average reading level of the retrieved web documents is relatively high. At the same time, documents at a wide range of reading levels are identified and even among the Top-10 search results one finds documents at the lower levels, supporting the potential usefulness of readability ranking for the web. Finally, we report on generalization experiments showing that the features we used generalize well across different web sources.

1 Introduction

The web is a vast source of information on a broad range of topics. While modern search engines make use of a range of features for identifying and ranking search results, the question whether a web page presents its information in a form that is accessible to a given reader is only starting to receive attention. Researching the use of readability assessment as a ranking parameter for web search can be a relevant step in that direction.

Readability assessment has a long history spanning various fields of research from Educational Psychology to Computer Science. At the same time, the question which features generalize to different types of documents and whether the readability models are appropriate for real-life applications has only received little attention.

Against this backdrop, we want to see how well a state-of-the-art readability assessment approach using a broad range of features performs when applied to web data. Based on the approach introduced in Vajjala and Meurers (2012), we thus set out to explore the following two questions in this paper:

- Which reading levels can be identified in a systematic sample of web texts?
- How well do the features used generalize to different web sources?

The paper is organized as follows: Section 2 surveys related work. Section 3 introduces the corpus and the features we used. Section 4 describes our readability models. Section 5 discusses our experiments investigating the applicability of these models to web texts. Section 6 reports on a second set of experiments conducted to test the generalizability of the features used. Section 7 concludes the paper with a discussion of our results.

2 Related Work

2.1 Readability Assessment

The need for assessing the readability of a piece of text has been explored in the educational research community for over eight decades. DuBay (2006) provides an overview of early readability formulae, which were based on relatively shallow features and wordlists. Some of the formulae are still being used in practice, as exemplified by the Flesch-Kincaid Grade Level (Kincaid et al., 1975) available in Microsoft Word.

More recent computational linguistic approaches view readability assessment as a

Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, pages 59–68,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

classification problem and explore different types of features. Statistical language modeling has been a popular approach (Si and Callan, 2001; Collins-Thompson and Callan, 2004), with the hypothesis that the word usage patterns across grade levels are distinctive enough. Heilman et. al. (2007; 2008) extended this approach by combining language models with manually and automatically extracted grammatical features.

The relation of text coherence and cohesion to readability is well explored in the CohMetrix project (McNamara et al., 2002). Ma et al. (2012a; 2012b) approached readability assessment as a ranking problem and also compared human versus automatic feature extraction for the task of labeling children's literature.

The WeeklyReader[1], an American educational newspaper with graded readers has been a popular source of data for readability classification research in the recent past. Petersen and Ostendorf (2009), Feng et al. (2009) and Feng (2010) used it to build readability models with a range of lexical, syntactic, language modeling and discourse features. In Vajjala and Meurers (2012) we created a larger corpus, *WeeBit*, by combining WeeklyReader with graded reading material from the BBCBitesize website.[2] We adapted measures of lexical richness and syntactic complexity from Second Language Acquisition (SLA) research as features for readability classification and showed that such measures of proficiency can successfully be used as features for readability assessment.

2.2 Readability Assessment of Web Texts

Despite the significant body of research on readability assessment, applying it to retrieve relevant texts from the web has elicited interest only in the recent past. While Bennöhr (2005) and Newbold et al. (2010) created new readability formulae for this purpose, Ott and Meurers (2010) and Tan et al. (2012) used existing readability formulae to filter search engine results. The READ-X project (Miltsakaki and Troutt, 2008; Miltsakaki, 2009) combined standard readability formulae with topic classification to retrieve relevant texts for users.

The REAP Project[3] supports the lexical acquisition of individual learners by retrieving texts that suit a given learner level. Kidwell et al. (2011) also used a word-acquisition model for readability prediction. Collins-Thompson et al. (2011) and Kim et al. (2012) employed word distribution based readability models for personalized search and for creating entity profiles respectively. Nakatani et al. (2010) followed a language modeling approach to rank search results to take user comprehension into account. Google also has an option to filter search results based on reading level, apparently using a language modeling approach.[4] Kanungo and Orr (2009) used search result snippet based features to predict the readability of short web-summaries.

All the above approaches primarily restrict themselves to traditional formulae or statistical language models encoding the distribution of words. The effect of lexical and syntactic features as used in recent research on readability thus remains to be studied in a web context. Furthermore, the generalizability of the features used to other data sets also remains to be explored. These are the primary issues we address in this paper.

3 Corpus and Features

Let us turn to answering our first question: Which reading levels can be identified in a systematic sample of web texts? To address this question, we first need to introduce the features we used, the graded corpus we used to train the model, and the nature of the readability model.

Since the goal of this paper is not to present new features but to explore the application of a readability approach to the web, we here simply adopt the feature and corpus setup introduced in Vajjala and Meurers (2012). The *WeeBit* corpus used is a corpus of texts belonging to five reading levels, corresponding to children of age group 7–16 years. It consists of 625 documents per reading level. The articles cover a range of fiction and non-fiction topics. Each article is labeled as belonging to one of five reading levels: Level 2, Level 3, Level 4, KS3 and GCSE.

We adapted both the lexical and syntactic features of Vajjala and Meurers (2012) to build readability models on the basis of the WeeBit corpus and then studied their applicability to real-world documents retrieved from the web as well as the applicability of those features across different web sources.

[1] http://weeklyreader.com
[2] http://www.bbc.co.uk/bitesize
[3] http://reap.cs.cmu.edu
[4] http://goo.gl/aVy93

Lexical features (LEXFEATURES) The lexical features are motivated by the lexical richness measures used to estimate the quality of language learners' oral narratives (Lu, 2012). We included several type-token ratio variants used in SLA research: generic type token ratio, root TTR, corrected TTR, bilogarithmic TTR and Uber Index.

In addition, there are lexical variation measures used to estimate the distribution of various parts of speech in the given text. They include the noun variation, adjective variation, modifier variation, adverb variation and verb variation, which represent the proportion of words of the respective part of speech categories compared to all lexical words in the document. Alternative measures for verb variation, namely, Squared Verb Variation and Corrected Verb Variation are also included. Apart from these, we also added the traditionally used measures of average number of characters per word, average number of syllables per word, and two readability formulae, the Flesch-Kincaid score (Kincaid et al., 1975) and the Coleman-Liau score (Coleman and Liau, 1975). Finally, we included the percentage of words from the Academic Word List[5]. It is a list created by Coxhead (2000) which consists of words that are more commonly found in academic texts.

Syntactic features (SYNFEATURES) These features are adapted from the syntactic complexity measures used to analyze second language writing (Lu, 2010). They are calculated based on the parser output of the BerkeleyParser (Petrov and Klein, 2007), using the Tregex (Levy and Andrew, 2006) pattern matcher. They include: mean lengths of various production units (sentence, clause and t-unit); clauses per sentence and t-unit; t-units per sentence; complex-t units per t-unit and per sentence; dependent clauses per clause, t-unit and sentence; co-ordinate phrases per clause, t-unit and sentence; complex nominals per clause and t-unit; noun phrases, verb phrases and preposition phrases per sentence; average length of NP, VP and PP; verb phrases per t-unit; SBARs per sentence and average parse tree height.

We refer to the feature subset containing all the traditionally used features (# char. per word, # syll. per word and # words per sentence) as TRADFEATURES in this paper.

[5] http://simple.wiktionary.org/wiki/Wiktionary:Academic_word_list

4 The Readability Model

In computational linguistics, readability assessment is generally approached as a classification problem. To our knowledge, only Heilman et al. (2008) and Ma et al. (2012a) experimented with other kinds of statistical models.

We approach readability assessment as a regression problem. This produces a model which provides a continuous estimate of the reading level, enabling us to see if there are documents that fall between two levels or above the maximal level found in the training data. We used the WEKA implementation of linear regression for this purpose. Since linear regression assumes that the data falls on an interval scale with evenly spaced reading levels, we used numeric values from 1–5 as reading levels instead of the original class names in the *WeeBit* corpus. Table 1 shows the mapping from WeeBit classes to numeric values, along with the age groups per class.

WeeBit class	Age (years)	Reading level
Level 2	7–8	1
Level 3	8–9	2
Level 4	9–10	3
KS3	11–14	4
GCSE	14–16	5

Table 1: *WeeBit* Reading Levels for Regression

We report Pearson's correlation coefficient and Root Mean Square Error (RMSE) as our evaluation metrics. Correlation coefficient measures the extent of linear relationship between two random variables. In readability assessment, a high correlation indicates that the texts at a higher difficulty level are more likely to receive a higher level prediction from the model and those at lower difficulty level would more likely receive a lower prediction. RMSE can be interpreted as the average deviation in grade levels between the predicted and the actual values.

We trained four regression models with the feature subsets introduced in section 3: LEXFEATURES, SYNFEATURES, TRADFEATURES and ALLFEATURES. While the criterion used in creating the graded texts in *WeeBit* is not known, it is likely that they were created with the traditional measures in mind. Indeed, the traditional features also were among the most predictive features in Vajjala and Meurers (2012). Hence, apart from

training the above mentioned four regression models, we also trained a fifth model excluding the traditional features and formulae. This experiment was performed to verify if the traditional features are creating a skewed model that relies too heavily on those well-known and thus easily manipulated features in making decisions on test data. We refer to this fifth feature group as NOTRAD.

Table 2 shows the result of our regression experiments using 10-fold cross-validation on the *WeeBit* corpus, employing the different feature subsets and the complete feature set.

Feature Set	# Features	Corr.	RMSE
LEXFEATURES	17	0.84	0.78
SYNFEATURES	25	0.88	0.64
TRADFEATURES	3	0.66	1.06
ALLFEATURES	42	**0.92**	**0.54**
NOTRAD	37	0.89	0.63

Table 2: Linear Regression Results for *WeeBit*

The best correlation of 0.92 was achieved with the complete feature set. 0.92 is considered a strong correlation and coupled with an RMSE of 0.54, we can conclude that our regression model is a good model. In comparison, in Vajjala and Meurers (2012), where we tackle readability assessment as a classification problem, we obtained 93.3% accuracy on this dataset using all features.

Looking at the feature subsets, there also is a good correlation between the model predictions and the actual results in the other cases, except for the model considering only traditional features. While traditional features often are among the most predictive features in readability research, we also found that a model which does not include them can perform at a comparable level (0.89).

Comparing these results with previous research using regression modeling for readability assessment is not particularly meaningful because of the differences in the corpus and the levels used. For example, while Heilman et al. (2008) used a corpus of 289 texts across 12 reading levels achieving a correlation of 0.77, we used the *WeeBit* corpus containing 3125 texts across 5 reading levels.[6]

We took the two best models of Table 2, MODALL using ALLFEATURES and MODNOTRAD using the NOTRAD feature set, and set out to answer our first guiding question, about the reading levels which such models can identify in a systematic sample of web texts.

[6]Direct comparisons on the same data set would be most indicative, but many datasets, such as the corpus used in Heilman et al. (2008), are not accessible due to copyright issues.

5 Applying readability models to web texts

To investigate the effect of the two readability models for real-world web texts, we studied their performance on two types of web data:

- web documents we crawled from specific web sites that offer the same type of material for two groups of readers differing in their reading skills
- web documents identified by a web search engine for a sample of web queries selected from a public query log

5.1 Readability of web data drawn from characteristic web sites

5.1.1 Web test sets used

Following the approach of Collins-Thompson and Callan (2005) and Sato et al. (2008), who evaluated readability models using independent web-based test sets, we compiled three sets of web documents that given their origin can be classified into two classes each:

Wiki – SimpleWiki: Wikipedia[7], along with its manually simplified version *Simple Wikipedia*[8] is increasingly used in two-class readability classification tasks and text simplification approaches (Napoles and Dredze, 2010; Zhu et al., 2010; Coster and Kauchak, 2011). We use a collection of 2000 randomly selected parallel articles from each of the two websites, which in the following is referred to as WIKI and SIMPLEWIKI.

Time – Time for Kids: *Time for Kids*[9] is a division of the TIME magazine[10], which produces articles exclusively for children and is used widely in classrooms. We took a sample of 2000 documents each from Time and from Time for Kids for our experiments and refer them TIME and TFK.

NormalNews – ChildrensNews: We crawled websites that contain news articles written for children (e.g., `http://www.firstnews.co.uk`) and categorized them as CHILDRENSNEWS. We also crawled freely accessible articles from popular news websites such as *BBC* or *The Guardian* and

[7]`http://en.wikipedia.org`
[8]`http://simple.wikipedia.org`
[9]`http://www.timeforkids.com`
[10]`http://www.time.com`

categorized them as NORMALNEWS. We took 10K documents from each of these two categories for our experiments.

These three corpus pairs collected as test cases differ in several aspects. For example, SimpleWikipedia is not targeting children as such, whereas Time for Kids and ChildrensNews are. And SimpleWikipedia – Wikipedia covers parallel articles in two versions, whereas this is not the case for the the two Time and the two News corpora. However, as far as we see these differences are orthogonal to the issue we are researching here, namely their use as real-life test cases to study the effect of the classification model learned on the WeeBit data.

We applied the two regression models which had performed best on the *WeeBit* corpus (cf. Table 2 in section 4) to these web datasets. The average reading levels of the different datasets according to these two models are reported in Table 3.

Data Set	MODALL	MODNOTRAD
SIMPLEWIKI	3.86	2.67
TFK	4.15	2.72
CHILDRENSNEWS	4.19	2.39
WIKI	4.21	3.33
TIME	5.04	4.07
NORMALNEWS	5.58	4.42

Table 3: Applying the *WeeBit* regression model to the six web datasets

The table shows that both MODALL and MODNOTRAD place the documents from the children websites (SIMPLEWIKI, TFK and CHILDRENSNEWS) at lower reading levels than those from the regular websites for adults (TIME, WIKI and NORMALNEWS). However, there is an interesting difference in the predictions made by the two models. The MODALL model including the traditional features consistently assigns a higher reading level to all the documents, and it also fails to separate CHILDRENSNEWS (4.19) from WIKI (4.20).

To be able to inspect this in detail, we plotted the class-wise reading level distribution of our regression models. Figure 1 shows the distribution of reading levels for these web datasets using MODALL. As we already knew from the averages, the model assigns somewhat higher reading levels to all documents, and the figure confirms that the texts for children (SIMPLEWIKI, TFK and CHILDRENSNEWS) are only marginally distinguished from the corresponding websites targeting adult readers (TIME,WIKI and NORMALNEWS). The NORMALNEWS dataset also seems to be placed in a much higher distribution compared to all the other test sets, with more than 50% of the documents getting a prediction of "higher" (the label used for documents placed at level 6 or higher).

Figure 2 shows the distribution of reading levels across the test sets according to MODNOTRAD, the model without traditional features. The model provides a broader coverage across all reading levels, with documents from children web sites and SimpleWikipedia clearly being placed at the lower end of the spectrum and web pages targeting adults at the higher end. NORMALNEWS documents are again placed the highest, but less than 10% fall outside the range established by WeeBit. TIME shows the highest diversity, with around 20% for each reading level above the lowest one.

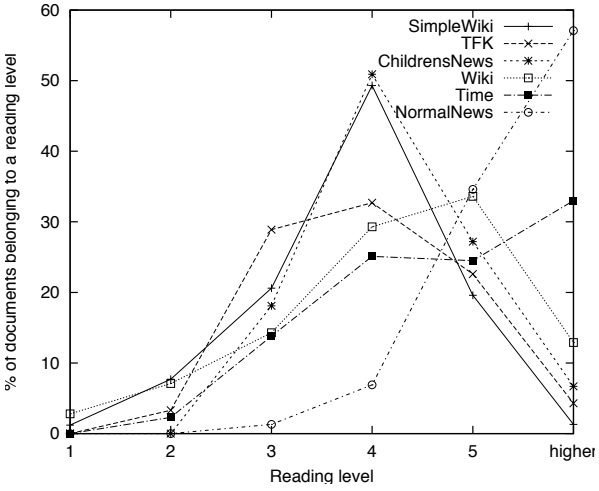

Figure 1: Reading levels assigned by MODALL

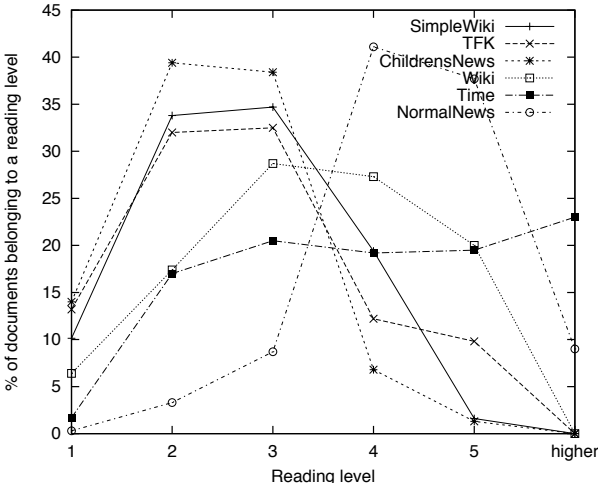

Figure 2: Reading levels using MODNOTRAD

The first set of experiments shows that the readability models which were successful on the *WeeBit* reference corpus seem to be able to identify a corresponding broad range among web documents that we selected top-down by relying on prototypical websites targeting "adult" and "child" readers, which are likely to feature more difficult and easier web documents, respectively. While we cannot evaluate the difference between the two models quantitatively, given the lack of an external gold standard classification of the crawled data, the MODNOTRAD conceptually seems to do a better job at distinguishing the two classes of websites in line with the top-down expectations.

5.2 Readability of search results

Complementing the first set of experiments, establishing that the readability models are capable of placing web documents in line with the top-down classification of the sites they originate from, in the second set of experiments we want to investigate bottom-up whether for some random topics of interest, the web offers texts at different readability levels. This also is of practical relevance, since ranking web search results by readability is only useful if there actually are documents at different reading levels for a given query.

For this investigation, we took the MODNOTRAD model and used it to estimate the reading level of web search results. For web searching, we used the BING search API (`http://datamarket.azure.com/dataset/bing/search`) and computed the reading levels of the Top-100 search results for a sample of 50 test queries, selected from a publicly accessible database (Lu and Callan, 2003).

Figure 3 characterizes the data obtained through the web searches in terms of the percentage of documents belonging to a given reading level, according to the MODNOTRAD model. In the Top-100 search results obtained for each of the 50 queries, the model identifies documents at all reading levels, with a peak at reading level 4 (corresponding to KS3 in the original WeeBit dataset).

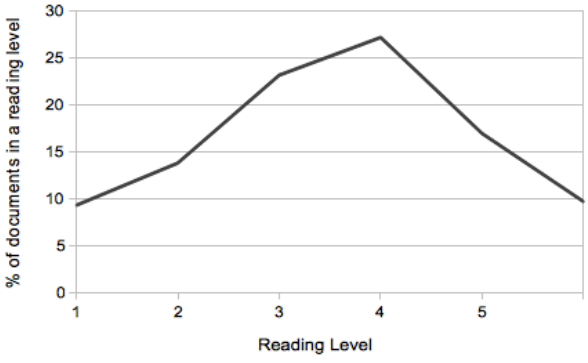

Figure 3: Documents retrieved per reading level

To determine how much individual queries differ in terms of the readability of the documents they retrieve, we also looked at the results for each query separately. Figure 4 shows the mean reading level of the Top-100 results for each of the 50 search queries. From query to query, the average readability of the documents retrieved seems to differ relatively little, with most results falling into the higher reading levels (4 or above).

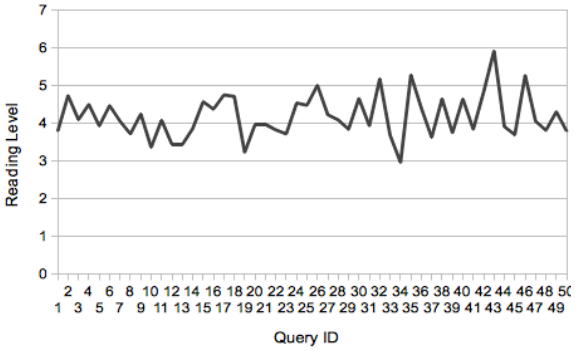

Figure 4: Average reading level of search results

Returning to the question whether there are documents of different reading levels for a given query, we need to check how much variation exists around the observed, rather similar averages. Table 4 provides the individual reading levels of the Top-10 search results for a sample of 10 queries from our experiment, along with the average reading level of the Top-100 results for that query. The results in Table 4 indicate that indeed there are documents at a broad range of reading levels even among the most relevant search results returned by the BING web search engine.

Looking at the individual query results, we found that although a lot of news documents tended towards a higher reading level, it is indeed possible to find some texts at lower reading levels even within Top-10 results (indicated in bold). However, we found that even for queries that we would expect to result in hits from websites targeting child readers, those sites often did not make it into the Top-10 results. The same was true for sites offering "simple" language, such as Simple Wikipedia, which was not among the top

Result Rank →	1	2	3	4	5	6	7	8	9	10	Avg$_{Top100}$
Query											
local anaesthetic	3.18	4.57	5.35	3.09	4.24	4.6	3.95	4.74	**2.72**	4.73	3.78
copyright copy law	1.77	4.59	1.43	2.67	4.63	6.2	2.69	**1.1**	3.87	5.61	4.57
halley comet	**1.69**	4.47	4.54	4.24	2.37	4.1	4.86	3.56	4.21	3.56	4.04
public offer	4.4	4.35	5.06	5.03	4.36	5.16	4.13	4.67	3.81	**1.1**	4.39
optic sensor	2.67	3.38	4.5	3.17	2.54	4.19	4.84	**1.47**	2.2	3.31	3.83
europe union politics	3.61	4.9	6.3	4.02	2.17	4.5	**1.47**	1.58	4.88	6.33	4.33
presidential poll	4.98	5.38	1.77	6.1	4.76	3.82	**1.05**	5.11	3.92	4.25	3.95
shakespeare	2.39	2.9	4.2	4.74	4.76	3.89	**1.47**	2.13	2.6	4.06	3.58
air pollution	**1.17**	4.93	3.7	2.3	4.36	3.73	3.71	3.49	2.22	2.67	4.21
euclidean geometry	3.88	4.71	4.7	4.3	4.45	4.63	4.04	4.1	3.48	**2.58**	3.18

Table 4: Reading levels of individual search results

results even when it contained pages directly relevant to the query. To provide access to those pages, reranking the search results based on readability would thus be of value.

While we do not want to jump to conclusions based on our sample of 50 queries, the results of our experiments seem to support the idea that readability-based re-ranking of web search results can help users in accessing web documents that also are at the right level for the given user. Returning to the first overall question that lead us here, our experiments support the answer that indeed there are documents spread across different reading levels on the web with a tendency towards higher reading levels.

6 Generalizability of the Feature Set

We can now turn to the second question raised in the introduction: How well do the features generalize across different classes of web documents? We saw in section 5.1 that the predictions of the two models we used varied quite a bit, solely based on whether the traditional readability features were included in the model or not. This confirms the need to investigate how generally applicable which types of features are across datasets.

As far as we know, such an experiment validating the generalizability of features was not yet performed in this domain. As there are no publicly available graded web datasets to build new readability models with the same feature set, we used the datasets we introduced in section 5.1.1 for creating two-class readability classification models. Since there are no clear age-group annotations with all these datasets, we decided to use a broad two-level classification instead of more fine grained grade levels.

The difference between this experiment and the previous one lies in the primary question it attempts to answer. Here, the focus is on verifying if the features are capable of building accurate classification models on different training sets. In the previous experiment, it was on checking if a given classification model (which in that experiment was trained on the WeeBit corpus) can successfully discriminate reading levels for documents from various real-world texts.

We observed in Section 5.1 that with traditional features, the WeeBit based readability model assigned higher reading levels to all the documents from our web datasets. So, it would perhaps be a natural step to train these binary classification models excluding the traditional features. However, the traditional features may still be useful (with different weights) for constructing classification models with other training data. So, we trained two sets of models per training set – one with ALLFEATURES and another excluding traditional features (NOTRAD).

We trained binary classification models using the following training sets:

- TIME – TFK texts
- WIKI – SIMPLEWIKI texts
- NORMALNEWS – KIDSNEWS texts
- TIME+WIKI – TFK+SIMPLEWIKI texts

We used the Sequential Minimal Optimization (SMO) algorithm implementation in the WEKA tool kit to train these classifiers. The choice of the algorithm here was motivated by the fact that training is quick and that SMO has successfully

been used in previous research on readability assessment (Feng, 2010; Hancke et al., 2012).

Table 5 summarizes the classification accuracies obtained with the four models using 10-fold cross validation for the four web corpora.

Training Set	Accuracy-All	Accuracy-NoTrad
TIME – TFK	95.11%	89.52%
WIKI – SIMPLEWIKI	92.32%	88.81%
NORMALNEWS – KIDSNEWS	97.93%	92.54%
TIME+WIKI – TFK+SIMPLEWIKI	93.38%	89.72%

Table 5: Cross-validation accuracies for binary classification on different web corpora

The results in the table show that the same set of features consistently result in creating accurate classification models for all four web corpora. Each of the two-class classification models performed well, despite the fact that the documents were created by different people and most likely with different instructions on how to write simple texts or simplify already existing texts. It was interesting to note the role of traditional features in improving the accuracy of these binary classification models. But, in the previous experiment, the model with traditional features consistently put all the documents into higher reading levels. It is possible that the role of traditional features in the WeeBit corpus may be skewed as it is likely that it was prepared with traditional readability measures in mind. Contrasting the results of these two experiments raises the question of what features hold more weight in what dataset, which is an interesting issue to explore in the future.

In sum, this experiment provides some clear evidence for affirmatively answering the second question about the generalizability of the feature set we used. The features seem to be sufficiently general for them to be useful in performing readability assessment of real-world documents.

7 Conclusion and Discussion

In this paper, we set out to investigate the applicability and generalizability of readability models for real-world web texts. We started with building readability models using linear regression, on a 5-level readability corpus with a range of lexical and syntactic features (section 4). We applied the two best models thus obtained to several web datasets we compiled from websites targeting children and others designed for adults (section 5.1) and on the Top-100 results obtained using a standard web search engine (section 5.2).

We observed that the models identified texts across a broad range of reading levels in the web corpora. Our pilot study of the reading levels of the search results confirmed that readability models could be useful as re-ranking or filtering parameters that prioritize relevant results which are at the right level for a given user. At the same time, we observed in both these experiments that the average reading level of general web articles is relatively high according to our models. Apart from result ranking, this also calls for the construction of efficient text simplification systems which pick up the difficult texts and attempt to simplify them to a given reading level.

We then proceeded to investigate how well the features used to build these readability models generalize across different corpora. For this, we reused the corpora with articles for children and adult readers from prototypical websites (section 5.1.1) and built four binary classification models with all of the readability features (section 6). Each of the models achieved good classification accuracies, supporting that the broad feature set used generalizes well across corpora. Whether or not to use traditional readability features is somewhat difficult to answer since those formulae are often taken into account when writing materials, so high classification accuracy on such corpora may be superficial in that it is not necessarily indicative of the spectrum of texts found on the web (section 5.1). This also raises the more general question which features work best for which kind of dataset. A systematic exploration of the effect of the individual features along with the impact of document topic and genre on readability would be interesting and relevant to pursue in the future.

In our future work, we also intend to explore further features for this task and improve our understanding of the correlations between the different features. Finally, we are considering reformulating readability assessment as ordinal regression or preference ranking.

Acknowledgements

We would like to thank the anonymous reviewers for their detailed, useful comments on the paper. This research was funded by the European Commission's 7th Framework Program under grant agreement number 238405 (CLARA).

References

Jasmine Bennöhr. 2005. A web-based personalised textfinder for language learners. Master's thesis, School of Informatics, University of Edinburgh.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Kevyn Collins-Thompson and Jamie Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL 2004*, Boston, USA.

Kevyn Collins-Thompson and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag. 2011. Personalizing web search results by reading level. In *Proceedings of the Twentieth ACM International Conference on Information and Knowledge Management (CIKM 2011)*.

William Coster and David Kauchak. 2011. Simple english wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA, June. Association for Computational Linguistics.

Averil Coxhead. 2000. A new academic word list. *Teachers of English to Speakers of Other Languages*, 34(2):213–238.

William H. DuBay. 2006. *The Classic Readability Studies*. Impact Information, Costa Mesa, California.

Lijun Feng, Nomie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 229–237, Athens, Greece, March. Association for Computational Linguistics.

Lijun Feng. 2010. *Automatic Readability Assessment*. Ph.D. thesis, City University of New York (CUNY).

Julia Hancke, Detmar Meurers, and Sowmya Vajjala. 2012. Readability classification for german using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1063–1080, Mumbay, India.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-07)*, pages 460–467, Rochester, New York.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications at ACL-08*, Columbus, Ohio.

Tapas Kanungo and David Orr. 2009. Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 202–211, New York, NY, USA. ACM.

P. Kidwell, G. Lebanon, and K. Collins-Thompson. 2011. Statistical estimation of word acquisition with application to readability prediction. In *Journal of the American Statistical Association. 106(493):21-30*.

Jin Young Kim, Kevyn Collins-Thompson, Paul N. Bennett, and Susan T. Dumais. 2012. Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 213–222, New York, NY, USA. ACM.

J. P. Kincaid, R. P. Jr. Fishburne, R. L. Rogers, and B. S Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. Research Branch Report 8-75, Naval Technical Training Command, Millington, TN.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Jie Lu and Jamie Callan. 2003. Content-based retrieval in hybrid peer-to-peer networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM'03)*.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Languages Journal*, pages 190–208.

Yi Ma, Eric Fosler-Lussier, and Robert Lofthus. 2012a. Ranking-based readability assessment for early primary children's literature. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 548–552, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yi Ma, Ritu Singh, Eric Fosler-Lussier, and Robert Lofthus. 2012b. Comparing human versus automatic feature extraction for fine-grained elementary readability assessment. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, PITR '12, pages 58–64, Stroudsburg, PA, USA. Association for Computational Linguistics.

Danielle S. McNamara, Max M. Louwerse, and Arthur C. Graesser. 2002. Coh-metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Proposal of Project funded by the Office of Educational Research and Improvement, Reading Program.

Eleni Miltsakaki and Audrey Troutt. 2008. Real time web text classification and analysis of reading difficulty. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*, pages 89–97, Columbus, Ohio. Association for Computational Linguistics.

Eleni Miltsakaki. 2009. Matching readers' preferences and reading skills with appropriate web texts. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, EACL '09, pages 49–52, Stroudsburg, PA, USA. Association for Computational Linguistics.

Makoto Nakatani, Adam Jatowt, and Katsumi Tanaka. 2010. Adaptive ranking of search results by considering user's comprehension. In *Proceedings of the 4th International Conference on Ubiquitous Information Management and Communication (ICUIMC 2010)*, pages 182–192. ACM Press, Suwon, Korea.

Courtney Napoles and Mark Dredze. 2010. Learning simple wikipedia: a cogitation in ascertaining abecedarian language. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, CL&W '10, pages 42–50, Stroudsburg, PA, USA. Association for Computational Linguistics.

Neil Newbold, Harry McLaughlin, and Lee Gillam. 2010. Rank by readability: Document weighting for information retrieval. In Hamish Cunningham, Allan Hanbury, and Stefan Rüger, editors, *Advances in Multidisciplinary Retrieval*, volume 6107 of *Lecture Notes in Computer Science*, pages 20–30. Springer Berlin / Heidelberg.

Niels Ott and Detmar Meurers. 2010. Information retrieval for education: Making search engines language aware. *Themes in Science and Technology Education. Special issue on computer-aided language analysis, teaching and learning: Approaches, perspectives and applications*, 3(1–2):9–30.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23:86–106.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April.

Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. 2008. Automatic assessment of japanese text readability based on a textbook corpus. In *LREC'08*.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*, pages 574–576. ACM.

Chenhao Tan, Evgeniy Gabrilovich, and Bo Pang. 2012. To each his own: Personalized content selection based on text comprehensibility. In *In Proceedings of WSDM*.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In Joel Tetreault, Jill Burstein, and Claudial Leacock, editors, *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7) at NAACL-HLT*, pages 163–173, Montral, Canada, June. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING), August 2010. Beijing, China*.