On The Applicability of Readability Models to Web Texts

Sowmya Vajjala and Detmar Meurers

University of Tübingen, Germany

PITR Workshop, ACL, 8th August 2013

CLARA

Outline of the Talk

- Background and Problem statement
- Experimental setup: Corpora, Features, Method
- Experiments
  1. Applying a state-of-the-art readability model to web texts
  2. Testing the generalizability of the feature set
- Conclusions and Discussion

Automatic Readability Assessment

- process of assessing the reading level of a text
  - potentially useful for humans as well as machines
- early research: readability formulae using sentence length, word length and lists of difficult words (cf. DuBay 2006)
- recent research:
  - language models (e.g., Collins-Thompson & Callan 2004; Schwarm & Ostendorf 2005)
  - syntactic parse features (e.g., Heilman et al. 2007)
  - cognitively motivated features (Feng et al. 2009)
  - coherence and cohesion (McNamara et al. 2002)
  - Second Language Acquisition based features (Vajjala & Meurers 2012)
  - ...

Applications of Readability Assessment

- filter search results through readability formulae (e.g., Bennöhr 2005; Ott & Meurers 2010; Tan et al. 2012)
- combine readability analysis with topic classification (Miltsakaki & Troutt 2008)
- personalized search (Collins-Thompson et al. 2011)
- for building user and topic profiles (Kim et al. 2012)

Our research questions

- Are state-of-the-art readability models actually useful for classifying texts as found on the web?
  - Which reading levels can be identified in a systematic sample of web texts?
  - Can re-ranking search results based on reading levels be useful?
- How well does the approach generalize?
  - the trained model?
  - the features?

Corpora: The WeeBit Corpus

- created for Vajjala & Meurers (2012)
- consists of articles belonging to 5 reading levels, for children from ages 7-16 years
- 625 documents per reading level
- We mapped the 5 reading levels to a scale of 1–5 to build a regression model.

Corpora: Two-Class Web Corpora

- We used other corpora for testing
  1. if WeeBit-trained regression model will identify the difference in reading levels between the two classes.
  2. if the feature set generalizes to other training sets.
- Two class corpora (difficult–easy) crawled from the web:
  - Wikipedia-Simple Wikipedia
  - Time-Time For Kids
  - Normal News-Children's News websites

Features

- We included the lexical and syntactic features from Vajjala & Meurers (2012).
- Lexical Features
  - lexical richness features from Second Language Acquisition (SLA) research
    - e.g., Type-Token ratio, noun variation, ...
  - POS density features
    - e.g., # nouns/# words, # adverbs/# words, ...
  - traditional features and formulae
    - e.g., # characters per word, Flesch-Kincaid, ...
- Syntactic Features
  - syntactic complexity features from SLA research.
    - e.g., # dep. clauses/clause, length of a t-unit, ...
  - other parse tree features
    - e.g., # NPs per sentence, avg. parse tree height, ...

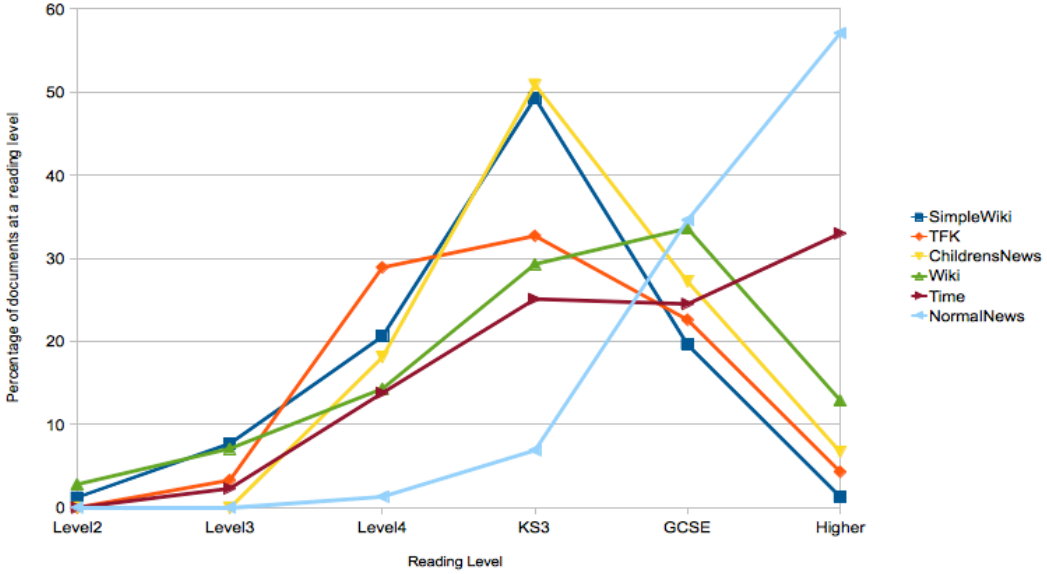
Approach

- Modeling: regression instead of classification
  - provides a continuous estimate on a scale
  - algorithm: linear regression (WEKA implementation)
- Evaluation:
  - measures: Pearson correlation, RMSE
  - method: 10 fold Cross Validation
- Model 1: with all features
  - Pearson corr. = 0.92, RMSE = 0.54
- Model 2: with all except traditional features
  - Pearson corr. = 0.89, RMSE = 0.63

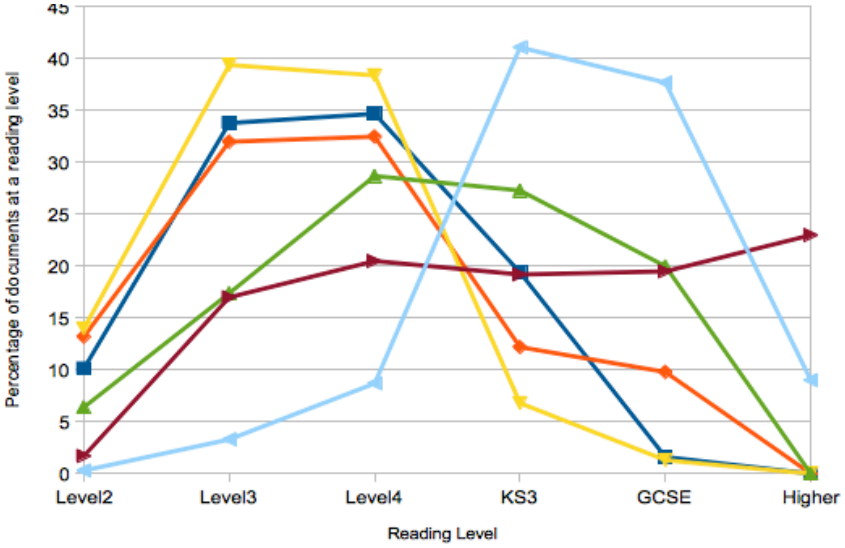
Applying Readability Models on Web Texts

Reading level distribution of web corpora

All Features



Without Traditional Features



- Notes:
  1. Regression values are rounded off to nearest integer and then mapped to original reading levels.
  2. ”Higher” is used for all values above the WeeBit scale.
- Is the model without traditional features better for assigning meaningful reading levels in the 1–5 range?

Applying Readability Models on Web Texts

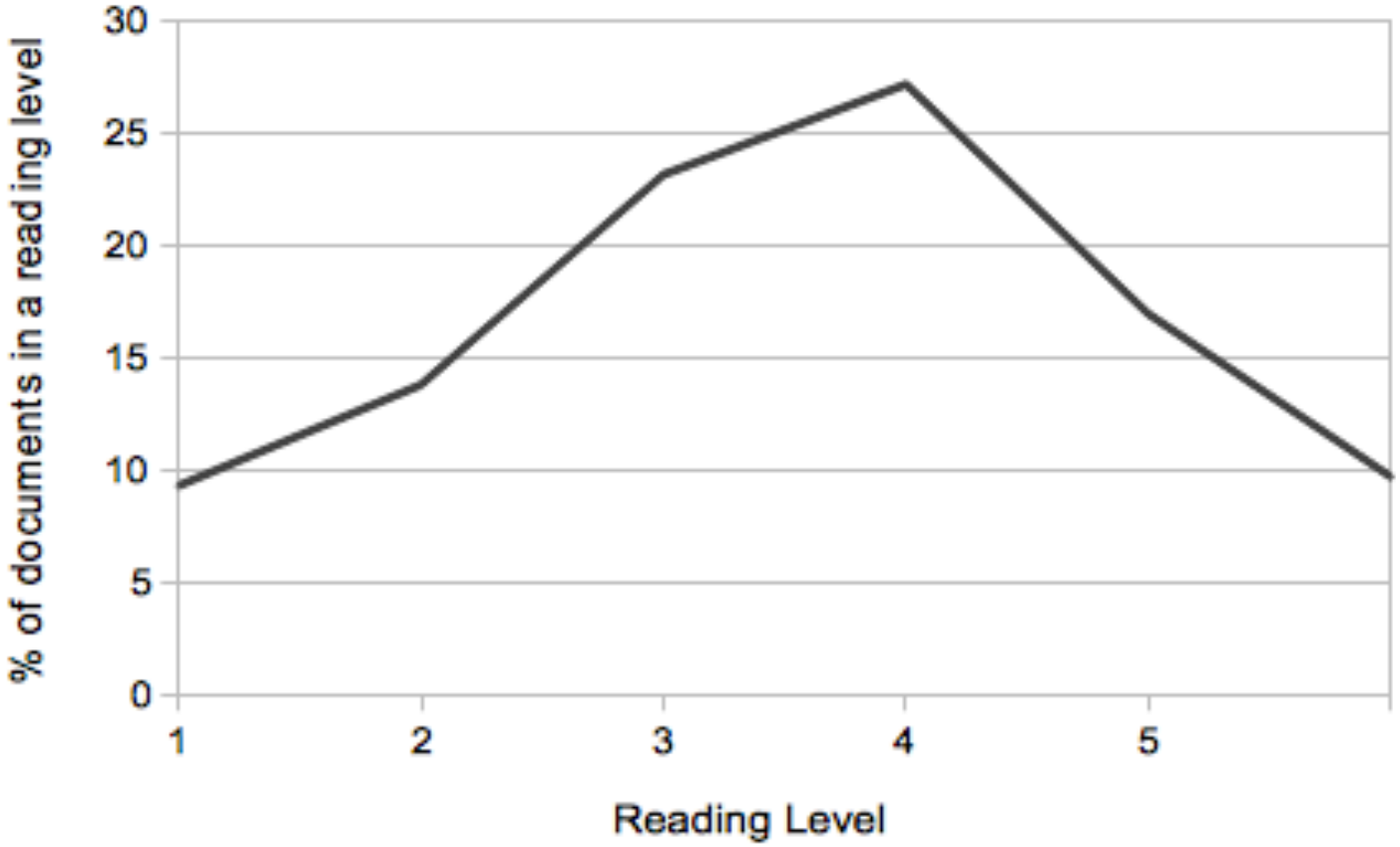
Reading levels of top search results

- Does the web offer texts at different reading levels for various topics of interest?
  - Can readability assessment be useful for re-ranking search engine result pages?
- We applied our readability model on search results obtained through BING search API.
  - took 50 search queries from public query log
  - computed reading levels for Top-100 results
  - results for example queries:

Result Rank →	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Avg**$_{Top100}$
Query											
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
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
shakespeare	2.39	2.9	4.2	4.74	4.76	3.89	**1.47**	2.13	2.6	4.06	3.58
euclidean geometry	3.88	4.71	4.7	4.3	4.45	4.63	4.04	4.1	3.48	**2.58**	3.18

Applying Readability Models on Web Texts

Reading level distribution of search results for 50 query sample



Applying Readability Models on Web Texts

What did we learn?

- Our results support the ideas that
  - our readability model can identify a broad range of reading levels across web texts.
  - readability based re-ranking of search results can be useful for identifying comprehensible results for users.
  - the average reading level of informative web texts is relatively high.
- Yet, the difference in predictions for the same text also raises the question of the generalizability of the features.

Generalizability of Features

Training and testing models for different web corpora

- How do we decide if the features are generalizable?
  - Train models for different corpora with the same feature set and test how accurate they are for binary classification.
- Results with 10 fold CV, using an SMO classifier:

Corpora	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%

Generalizability of Features

What did we learn?

- The features generalize well to other corpora, building successful readability classification models.
- Traditional features improve the classification accuracy, when training and testing on the same type of data.
- Which features are most informative in which domains?
  - Where does a single trained model generalize well and when is retraining with the feature set needed?

Summary

- Our readability model identified texts across a broad range of reading levels in the web corpora.
- A pilot study of reading levels of search results confirmed that:
  - Readability based re-ranking of search results is useful even amongst the most relevant results.
  - However, average reading level is high among the top-100 results.

  → Good motivation for text simplification
- The features generalized well across different corpora building efficient readability models over them.
  - traditional features seem to be are more important in specially prepared texts compared to general web texts.

Future Work

- Systematically explore the predictive power of individual features across datasets.
- Study the impact of topic and genre on readability.
- Explore the utility of other features.
- Study the correlations between features.
- Investigate a reformulation of readability assessment as ordinal regression or preference ranking.

Thank you!

Thank you for your attention
Questions? :-)

References

Bennöhr, J. (2005). A Web-based Personalised Textfinder for Language Learners. Master's thesis, School of Informatics, University of Edinburgh.

Collins-Thompson, K., P. N. Bennett, R. W. White, S. de la Chica & 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)*.

Collins-Thompson, K. & J. Callan (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL 2004*. Boston, USA. URL http://www.cs.cmu.edu/~callan/Papers/hlt04-kct.pdf.

DuBay, W. H. (2006). *The Classic Readability Studies*. Costa Mesa, California: Impact Information.

Feng, L., N. Elhadad & M. Huenerfauth (2009). Cognitively Motivated Features for Readability Assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, Greece: Association for Computational Linguistics, pp. 229–237. URL http://aclweb.org/anthology/E09-1027.

Heilman, M., K. Collins-Thompson, J. Callan & M. 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)*. Rochester, New York, pp. 460–467.

Kim, J. Y., K. Collins-Thompson, P. N. Bennett & S. 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*. New York, NY, USA: ACM, WSDM '12, pp. 213–222. URL http://doi.acm.org/10.1145/2124295.2124323.

McNamara, D. S., M. M. Louwerse & A. 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. URL http://cohmetrix.memphis.edu/cohmetrixpr/archive/Coh-MetrixGrant.pdf.

Miltsakaki, E. & A. 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*. Columbus, Ohio: Association for Computational Linguistics, pp. 89–97. URL http://aclweb.org/anthology/W08-0911.

Ott, N. & D. 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. URL http://purl.org/dm/papers/ott-meurers-10.html.

Schwarm, S. & M. Ostendorf (2005). Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*. Ann Arbor, Michigan, pp. 523–530.

Tan, C., E. Gabrilovich & B. Pang (2012). To Each His Own: Personalized Content Selection based on Text Comprehensibility. In *In Proceedings of WSDM*.

Vajjala, S. & D. Meurers (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In J. Tetreault, J. Burstein & C. Leacock (eds.), *In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*. Montréal, Canada: Association for Computational Linguistics, pp. 163—-173. URL http://aclweb.org/anthology/W12-2019.pdf.