On The Applicability of Readability Models to Web Texts

Sowmya Vajjala and Detmar Meurers
Department of Linguistics, University of Tübingen

In a Nutshell

We built a readability model using lexical and syntactic features and used it to answer the following questions:

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

Background

- Automatic readability assessment is a well studied problem.
  - early research: readability formulae using surface features (e.g., Kincaid et al., 1975)
  - recent research: machine learning based classification models (e.g., Feng et al., 2009; Vajjala & Meurers, 2012)
- Applications of readability assessment were primarily studied in the context of filtering search results. (e.g., Kim et al., 2012)
- But are readability models useful for classifying web texts into a broad range of reading levels?

  ⇒ We explore this question here.

Corpora and Features used

- **Corpora used**:
  - **WeeBit Corpus** (Vajjala & Meurers, 2012)
    - ∗ our primary corpus
    - ∗ consists of five reading levels, 625 articles per level
    - ∗ converted five levels to a scale of 1-5 (classification to regression)
  - **Two-class Corpora: Easy–Difficult**
    - ∗ Simple Wiki-Wikipedia (2000 pairs of parallel articles)
    - ∗ Time for Kids - Time (2000 articles per category)
    - ∗ Childrens News - Normal News (10K per category)

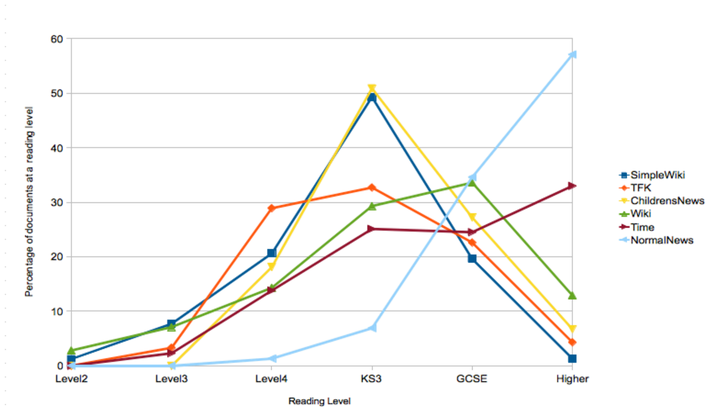
    Used to test the primary readability model and to build binary classification models.
- Features, adapted from Vajjala & Meurers (2012)
  - Lexical Features
    - ∗ features from Second Language Acquisition (SLA) research, POS densities, and traditional features
  - Syntactic Features
    - ∗ syntactic complexity features from SLA research, other parse tree features

Experimental setup

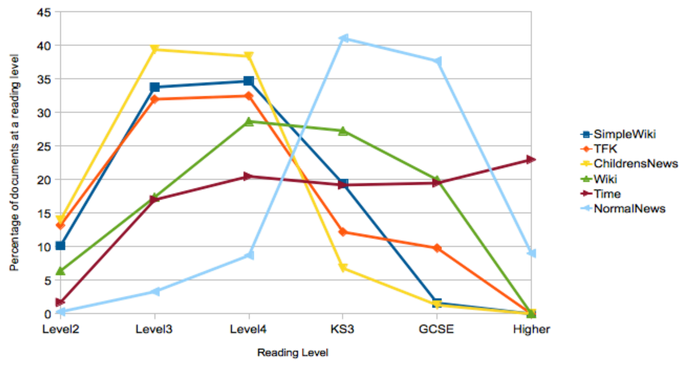
- Readability assessment as regression (not classification)
  - provides a continuous estimate
- Algorithm: linear regression
  (no significant difference for other regression options)
- Evaluation measures: Pearson correlation, RMSE
- Results for the models
  1. with all features: Pearson = 0.92, RMSE=0.54
  2. without traditional features: Pearson = 0.89, RMSE =0.63

Readability of web texts

- Model with all features



- Model without traditional features

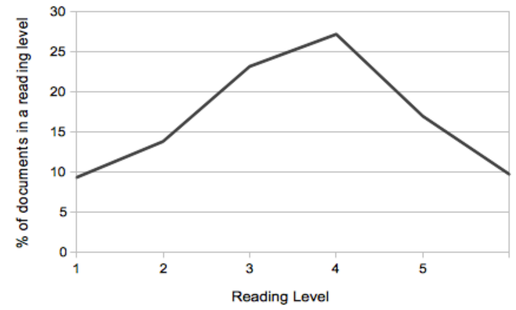


- What can we infer?
  - The models successfully identify differences in reading levels of web texts.
  - The model without traditional features seem to identify a broader range of reading levels.

Readability of search results

- Can readability assessment be useful for search engines?
  - We applied one of our readability models (the one without traditional features) to search results from BING.
  - Example reading levels of Top-10 results (50 queries):

Result Rank →	1	2	3	4	5	6	7	8	9	10	Avg_{Top100}
Query											
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



- What can we infer?
  - Readability-based search result re-ranking can be useful.
  - Average reading level of search results is relatively high.

Generalizability of features

- Will the features generalize to different corpora?

Training Set	All features	No Trad.
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%

⇒ Yes, the features generalize well across corpora.

- Note that traditional features work well here.
  Which features are most useful for which corpora?

Conclusions

- Our readability models are useful to identify a broad range of reading levels in web texts and search results.
- Average reading level of web texts is relatively high, which calls for the development of good text simplification systems.
- Our features generalize well across different web sources.
- Future work: Explore which features are important for which corpora, and understand correlations between them.