On The Applicability of Readability Models to Web Texts

Sowmya Vajjala and Detmar Meurers
University of Tübingen, Germany

PITR Workshop, ACL, 8th August 2013
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)
  - ...

On The Applicability of Readability Models to Web Texts
Sowmya Vajjala and Detmar Meurers
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:
  - 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:

<table>
<thead>
<tr>
<th>Result Rank →</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Avg Top100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>copyright copy law</td>
<td>1.77</td>
<td>4.59</td>
<td>1.43</td>
<td>2.67</td>
<td>4.63</td>
<td>6.2</td>
<td>2.69</td>
<td>1.1</td>
<td>3.87</td>
<td>5.61</td>
<td>4.57</td>
</tr>
<tr>
<td>halley comet</td>
<td>1.69</td>
<td>4.47</td>
<td>4.54</td>
<td>4.24</td>
<td>2.37</td>
<td>4.1</td>
<td>4.86</td>
<td>3.56</td>
<td>4.21</td>
<td>3.56</td>
<td>4.04</td>
</tr>
<tr>
<td>europe union politics</td>
<td>3.61</td>
<td>4.9</td>
<td>6.3</td>
<td>4.02</td>
<td>2.17</td>
<td>4.5</td>
<td>1.47</td>
<td>1.58</td>
<td>4.88</td>
<td>6.33</td>
<td>4.33</td>
</tr>
<tr>
<td>shakespeare</td>
<td>2.39</td>
<td>2.9</td>
<td>4.2</td>
<td>4.74</td>
<td>4.76</td>
<td>3.89</td>
<td>1.47</td>
<td>2.13</td>
<td>2.6</td>
<td>4.06</td>
<td>3.58</td>
</tr>
<tr>
<td>euclidean geometry</td>
<td>3.88</td>
<td>4.71</td>
<td>4.7</td>
<td>4.3</td>
<td>4.45</td>
<td>4.63</td>
<td>4.04</td>
<td>4.1</td>
<td>3.48</td>
<td><strong>2.58</strong></td>
<td>3.18</td>
</tr>
</tbody>
</table>
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.
On The Applicability of Readability Models to Web Texts
Sowmya Vajjala and Detmar Meurers

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:

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Accuracy-All</th>
<th>Accuracy-NoTrad</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIME – TFK</td>
<td>95.11%</td>
<td>89.52%</td>
</tr>
<tr>
<td>WIKI – SIMPLEWIKI</td>
<td>92.32%</td>
<td>88.81%</td>
</tr>
<tr>
<td>NORMALNEWS – KIDSNEWS</td>
<td>97.93%</td>
<td>92.54%</td>
</tr>
<tr>
<td>TIME+WIKI – TFK+SIMPLEWIKI</td>
<td>93.38%</td>
<td>89.72%</td>
</tr>
</tbody>
</table>
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 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!
On The Applicability of Readability Models to Web Texts
Sowmya Vajjala and Detmar Meurers

Introduction
Our Research Questions
Setup
Corpora
Features
Approach
Experiment 1
Reading levels in web corpora
Reading levels of pages obtained by web searches
Experiment 2
Generalizability of features
Summary
Future Work

References


On The Applicability of Readability Models to Web Texts
Sowmya Vajjala and Detmar Meurers

Introduction

Our Research Questions

Setup

Corpora

Features

Approach

Experiment 1

Reading levels in web corpora

Reading levels of pages obtained by web searches

Experiment 2

Generalizability of features

Summary

Future Work


On The Applicability of Readability Models to Web Texts
Sowmya Vajjala and Detmar Meurers

Introduction
Our Research Questions

Setup
Corpora
Features
Approach

Experiment 1
Reading levels in web corpora
Reading levels of pages obtained by web searches

Experiment 2
Generalizability of features

Summary

Future Work