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):

<table>
<thead>
<tr>
<th>Query</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</th>
</tr>
</thead>
<tbody>
<tr>
<td>halley comet</td>
<td>1.69</td>
<td>4.47</td>
<td>4.54</td>
<td>4.24</td>
<td>2.97</td>
<td>4.1</td>
<td>4.86</td>
<td>3.56</td>
<td>4.21</td>
<td>5.56</td>
<td>4.94</td>
</tr>
<tr>
<td>europe union politics</td>
<td>3.61</td>
<td>4.5</td>
<td>4.3</td>
<td>4.12</td>
<td>3.17</td>
<td>4.5</td>
<td>4.87</td>
<td>1.58</td>
<td>4.08</td>
<td>6.33</td>
<td>4.33</td>
</tr>
</tbody>
</table>

- 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?
  ⇒ 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.