Assessing the Relative Reading level of Sentence Pairs for Text Simplification

Sowmya Vajjala and Detmar Meurers
LEAD Graduate School and Department of Linguistics, University of Tübingen

In a Nutshell

Our Questions
- Can readability models be effective at the sentence level?
- Can we evaluate simplification using readability models?

Practical Relevance
- identify targets for text simplification
- evaluate simplification approaches

Approach

- Build a document-level model and evaluate it using a standard dataset.
- Apply the model to sentences using normal and simplified Wikipedia.

Corpora
1. (a) Document-level model: WeeBit corpus (Vajjala & Meurers, 2012)
2. Standard dataset for testing: Common Core Standards corpus
3. Sentence level: Wiki (N)-SimpleWiki(S) sentence aligned corpus (Zhu et al., 2010)

Modeling
- We model readability as regression (1–5).
- Evaluation of the document-level model:
  - Pearson Correlation (r) and Root Mean Square Error (RMSE)
- Evaluation of the model for sentences:
  - How well does the model identify S<N in the Wiki corpus?

Features
- Lexical Features
  - lexical richness features
    - TTR, noun variation, …
  - POS density features
    - # nouns/# words, …
  - traditional features and formulae
    - Flesch-Kincaid score, …
- Syntactic Features
  - complexity features from SLA
    - # dep. clauses/clause, …
  - other parse tree features
    - # NPs per sentence, …
  - Morphological properties of words
  - Avg. # of senses per word (WordNet)
  - Psycholinguistic features
    - Average Age-of-Acquisition of words (Kuperman et al., 2012)
    - word abstractness, …

Document Level Model
- Machine learner: SMOReg (WEKA)
- Result (WeeBit, 10-fold cross-validation)
  - r = 0.92, RMSE = 0.53
- Result on Common Core Standards

<table>
<thead>
<tr>
<th>Our System</th>
<th>Spearman</th>
<th>Pearson r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nelson et al. (2012):</td>
<td>0.69</td>
<td>0.61</td>
</tr>
<tr>
<td>REAP</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>ATOS</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>DRP</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>Lexile</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>Reading Maturity</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>SourceRater</td>
<td>0.75</td>
<td></td>
</tr>
</tbody>
</table>

Sentence-Level Model
- We started with training a sentence-level readability model on Wiki-SimpleWiki:
  - Binary classification: simple – hard
    - 65–68% accuracy, irrespective of the training sample size
  - As regression: \( r = 0.4 \)
- Poor performance raises two questions:
  1. Is simplification better understood as relative (instead of absolute)?
  2. Which features work best at sentence level?

Is Simplification Relative?
- How can we show this?
  - compute reading levels of normal (N) and simplified (S) sentences
  - evaluate classification using \( S < N \), \( S = N \) and \( S > N \)
  - How big must \( S − N \) be to interpret it as a categorical difference in reading level?
  \( \Rightarrow \) We call this the d-value.

Influence of N
- How does the reading level of the unsimplified sentence (N) affect the results?

Results for harder sentences (N > 2.5):

\[ \text{corpus (Vajjala & Meurers, 2012)} \]

Results for simpler sentences (N<2.5):

\[ \Rightarrow \text{It is difficult to identify simplifications for an already simple sentence.} \]

\[ \Rightarrow \text{Works well for more complex sentences.} \]

Summary
- Relative analysis of readability better for sentence-level comparison.
- The accuracy of comparison depends on:
  1. minimum \( S − N \) required to identify a categorical difference \( d \).
  2. reading level of the original, unsimplified sentence \( N \).
- Our approach works well for complex sentences.

Future Work
- Explore automatic identification of transformations for simplification.
- Explore ranking or ordinal regression instead of linear regression.

Influence of d

References