Automatic Readability Assessment
Features, models and their applicability

Sowmya Vajjala
with Detmar Meurers and Julia Hancke

Center for Natural Language Processing
Université Catholique de Louvain, Belgium
28th March 2014
Outline of the Talk

▶ Background
  ▶ What is readability assessment?
  ▶ Where is it useful?
  ▶ How do we build an automatic model for readability?

▶ Our Approach
  ▶ Corpus, Features and Modeling

▶ Applicability to real world
  ▶ for various domains and genres
  ▶ sentences instead of documents
  ▶ practical applications

▶ Readability Assessment for German
What is readability assessment?

Measuring **how difficult** it is to read a text,

- given a **purpose**, e.g.,
  - identifying appropriate texts for the readers
  - evaluation of natural language generation systems

- based on **properties of the text** using criteria which are
  - theory-driven (e.g., difficult syntactic constructions)
  - data-induced (e.g., corpora with graded texts)
  - in-between (e.g., derived frequency information for words)

- for a **user**, given some information about him
  (e.g., language ability, age, working memory)
  - obtained directly (e.g., questionnaire), or
  - indirectly (e.g., inferred from nature of a search query)
Where is this useful?

▶ retrieving relevant reading materials for language learners (e.g., REAP project, http://reap.cs.cmu.edu)
▶ evaluating the quality of abstracts and automatically generated summaries (Kanungo & Orr 2009)
▶ readers with special needs (e.g., adults with intellectual disabilities as in Huenerfauth et al. 2009)
▶ personalized search (Liu et al. 2004)
▶ assessing difficulty of questions in survey questionnaires (?)
Where is this useful?

Our work

- text simplification (Vajjala & Meurers, EACL 2014)
  - application context: bilingual classrooms (project starting in June 2014)
- identifying television programs into age-groups based on subtitles (Vajjala & Meurers, PITR 2014)
- comparing the linguistic complexity of German text books across school types & grades (ongoing project)
- analyzing impact of linguistic complexity of questionnaire data on the results (Göllner et al, GEBF 2014 talk)
How is readability assessment done?

Traditional Approaches

- *Traditional readability formulae* used shallow, easy to quantify features
  - e.g., average sentence length, average word length
  - Kincaid et al. (1975), DuBay (2004), . . .

- Some others were based on *lexical measures*
  - e.g., words on specific word lists

- These approaches are popular even now (e.g., in Word).

- Tools exist for calculating scores using
  - popular formulae (e.g., https://readability-score.com)
  - proprietary approaches (e.g., http://www.lexile.com)
How is readability assessment done?

Computational Approaches

- **Language n-gram models**
  (Si & Callan 2001; Collins-Thompson & Callan 2004; Petersen & Ostendorf 2009)

- **Machine learning using lexical and syntactic features**
  (Schwarm & Ostendorf 2005; Heilman et al. 2007; Feng 2010)

- **Language specific morphological features**
  (Dell’Orletta et al. 2011; Francois & Fairon 2012; Hancke et al. 2012)

- **Features motivated by cognitive perspective**
  (Crossley & McNamara 2011)

- **Features motivated by Second Language Acquisition (SLA)**
  - measures of language development or proficiency can successfully be used as measures of text complexity
    (Vajjala & Meurers 2012)
Typical Steps in a Machine Learning Approach

- You have some texts annotated with reading level (corpus) or you collect some graded texts.
- Start with a hypothesis about text complexity and quantify the hypothesis (features).
- How will you build a machine learning model? (modeling)
  - Group it in to discrete readability levels (= classification)
    e.g., beginner, intermediate, advanced
  - continuous readability levels (= regression)
    e.g., grade 1, 2, ... 
- Verify the performance (evaluation) and use this model for predicting the reading level of new texts.
Our Approach: Overview

- **Corpus**: accessible graded corpora (not created by us)
- **Features**: utilize insights on linguistic complexity from Second Language Acquisition and Psycholinguistics
- **Modeling and Evaluation**: build and evaluate regression and classification models
- **Real-life Evaluation**: test applicability by performing multiple evaluations with real-world test sets
The WeeBit Corpus (Vajjala & Meurers 2012)

We combined two reading level annotated corpora to create a new corpus: WeeBit.

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>Age in Years</th>
<th>Number of Articles</th>
<th>Avg. Number of Sentences/Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>from WeeklyReader</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>7–8</td>
<td>629</td>
<td>23.41</td>
</tr>
<tr>
<td>Level 3</td>
<td>8–9</td>
<td>801</td>
<td>23.28</td>
</tr>
<tr>
<td>Level 4</td>
<td>9–10</td>
<td>814</td>
<td>28.12</td>
</tr>
<tr>
<td>from BBCBitesize</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KS3</td>
<td>11–14</td>
<td>644</td>
<td>22.71</td>
</tr>
<tr>
<td>GCSE</td>
<td>14–16</td>
<td>3500</td>
<td>27.85</td>
</tr>
</tbody>
</table>

Note: If you want to use this corpus for your experiments, please contact me after the talk or send an email later!
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 score, ...

- **Syntactic Features**
  - syntactic complexity features from SLA research.
    - e.g., # dep. clauses/clause, average clause length, ...
  - other parse tree features
    - e.g., # NPs per sentence, avg. parse tree height, ...
Features from Vajjala & Meurers (2014a)

- Morphological properties of words
  - e.g., Does the word contain a stem along with an affix? *abundant* = *abound* + *ant*

- Age of Acquisition (AoA)
  - average age-of-acquisition of words in a text

- Other Psycholinguistic features
  - e.g., word abstractness

- Avg. number of senses per word (obtained from WordNet)
Implementation details
Tools, Resources and Algorithms used

▶ Tools:
  ▶ For Lexical Features
    ▶ Stanford Tagger (Toutanova, Klein, Manning & Singer 2003)
  ▶ For Syntactic Features
    ▶ Berkeley Parser (Petrov & Klein 2007)
    ▶ Tregex Pattern Matcher (Levy & Andrew 2006)
  ▶ For Machine Learning algorithms
    ▶ WEKA (http://www.cs.waikato.ac.nz/ml/weka/)
    ▶ modeled readability as regression, using SMOReg algorithm.

▶ Resources:
  ▶ Celex Lexical Database (http://celex.mpi.nl)
  ▶ Kuperman et al. (2012)’s AoA ratings
  ▶ MRC Psycholinguistic database (http://ota.oucs.ox.ac.uk/headers/1054.xml)
  ▶ Wordnet Database (http://wordnet.princeton.edu)
How good is our model?

- performance with 10-fold Cross-validation
  - Pearson Correlation = 0.9
  - Room Mean Square Error (RMSE) = 0.53
How good is our model?

- performance with 10-fold Cross-validation
  - Pearson Correlation = 0.9
  - Room Mean Square Error (RMSE) = 0.53
- Performance on a standard test set
  - Common Core Standards example set
  - 168 texts belonging to grade levels 2–12
  - Evaluation: Spearman’s rank correlation (scales differ)
How good is our model?

- performance with 10-fold Cross-validation
  - Pearson Correlation = 0.9
  - Room Mean Square Error (RMSE) = 0.53
- Performance on a standard test set
  - Common Core Standards example set
  - 168 texts belonging to grade levels 2–12
  - Evaluation: Spearman’s rank correlation (scales differ)

<table>
<thead>
<tr>
<th>System</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our System</td>
<td>0.69</td>
</tr>
<tr>
<td>Nelson et al. (2012):</td>
<td></td>
</tr>
<tr>
<td>REAP (<a href="http://reap.cs.cmu.edu">http://reap.cs.cmu.edu</a>)</td>
<td>0.54</td>
</tr>
<tr>
<td>ATOS (<a href="http://renlearn.com/atos">http://renlearn.com/atos</a>)</td>
<td>0.59</td>
</tr>
<tr>
<td>DRP (<a href="http://questarai.com/Products/DRPProgram">http://questarai.com/Products/DRPProgram</a>)</td>
<td>0.53</td>
</tr>
<tr>
<td>Lexile (<a href="http://lexile.com">http://lexile.com</a>)</td>
<td>0.50</td>
</tr>
<tr>
<td>Reading Maturity (<a href="http://readingmaturity.com">http://readingmaturity.com</a>)</td>
<td>0.69</td>
</tr>
<tr>
<td>SourceRater (<a href="http://naeptba.ets.org/SourceRater3">http://naeptba.ets.org/SourceRater3</a>)</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Does this approach work on real-world texts? (Vajjala & Meurers 2014c)

- How well does our readability model generalize?
- Do our features generalize for other datasets?
- Is there a topic or genre effect?
- What happens if we analyze sentences instead of texts?
  → How accurately can we identify the differences in the sentential reading level before and after simplification?
Does this approach work on real-world texts? (Vajjala & Meurers 2014c)

- How well does our readability model generalize?
- Do our features generalize for other datasets?
- Is there a topic or genre effect?
- What happens if we analyze sentences instead of texts?
How does our model generalize?  
(Vajjala & Meurers 2014c)

- How will we know?
  - test the model with other reading level annotated datasets.
  - created by different groups, with different purposes
How does our model generalize?
(Vajjala & Meurers 2014c)

How will we know?

- test the model with other reading level annotated datasets.
- created by different groups, with different purposes

<table>
<thead>
<tr>
<th>Test set</th>
<th>Description</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>CommonCore</td>
<td>168 docs, scale: 2-12</td>
<td>0.69</td>
</tr>
<tr>
<td>TASA corpus</td>
<td>37K docs, scale: 28-110</td>
<td>0.86</td>
</tr>
<tr>
<td>Math Readability Corpus</td>
<td>120 docs, scale: 1-7</td>
<td>0.29</td>
</tr>
</tbody>
</table>
Does this approach work on real-world texts? (Vajjala & Meurers 2014c)

- How well does our readability model generalize?
- Do our features generalize for other datasets?
- Is there a topic or genre effect?
- What happens if we analyze sentences instead of texts?
How do the features generalize? (Vajjala & Meurers 2014c)

- How will we know?
  - We trained regression models with multiple datasets.
How do the features generalize?
(Vajjala & Meurers 2014c)

- How will we know?
  - We trained regression models with multiple datasets.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Description</th>
<th>Pearson</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>WeeBit</td>
<td>625 docs per level, 5 levels</td>
<td>0.92</td>
<td>0.53</td>
</tr>
<tr>
<td>CommonCore</td>
<td>168 docs, scale: 2-12</td>
<td>0.59</td>
<td>2.69</td>
</tr>
<tr>
<td>TASA</td>
<td>37K docs, scale: 28-110</td>
<td>0.97</td>
<td>1.77</td>
</tr>
<tr>
<td>Math Readability</td>
<td>120 docs, scale: 1-7</td>
<td>0.51</td>
<td>1.73</td>
</tr>
</tbody>
</table>
Does this approach work on real-world texts? (Vajjala & Meurers 2014c)

- How well does our readability model generalize?
- Do these features generalize for other datasets?
- Is there a topic or genre effect?
- What happens if we analyze sentences instead of texts?
Is there a topic or genre effect?  
(Vajjala & Meurers 2014c)
Is there a topic or genre effect? (Vajjala & Meurers 2014c)

- How will we know?
  - Common Core standards test set has genre annotation.

(Vajjala & Meurers 2014c)
Is there a topic or genre effect? (Vajjala & Meurers 2014c)

- How will we know?
  - Common Core standards test set has genre annotation.

- Genre-wise performance of our model on the dataset:

<table>
<thead>
<tr>
<th>Genre</th>
<th>Spearman’s Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>0.35</td>
</tr>
<tr>
<td>Literature</td>
<td>0.51</td>
</tr>
<tr>
<td>Informative</td>
<td>0.76</td>
</tr>
<tr>
<td>Misc.</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Is there a topic or genre effect?
(Vajjala & Meurers 2014c)

- How will we know?
  - Common Core standards test set has genre annotation.

- Genre-wise performance of our model on the dataset:

<table>
<thead>
<tr>
<th>Genre</th>
<th>Spearman’s Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>0.35</td>
</tr>
<tr>
<td>Literature</td>
<td>0.51</td>
</tr>
<tr>
<td>Informative</td>
<td>0.76</td>
</tr>
<tr>
<td>Misc.</td>
<td>0.69</td>
</tr>
</tbody>
</table>

- Can this effect be overcome?

"Yes, by training genre-specific models." (Vajjala & Meurers 2014b)
Is there a topic or genre effect? (Vajjala & Meurers 2014c)

- How will we know?
  - Common Core standards test set has genre annotation.

- Genre-wise performance of our model on the dataset:

<table>
<thead>
<tr>
<th>Genre</th>
<th>Spearman’s Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>0.35</td>
</tr>
<tr>
<td>Literature</td>
<td>0.51</td>
</tr>
<tr>
<td>Informative</td>
<td>0.76</td>
</tr>
<tr>
<td>Misc.</td>
<td>0.69</td>
</tr>
</tbody>
</table>

- Can this effect be overcome?
  - Yes, by training genre-specific models.
  - Vajjala & Meurers (2014b)
Tackling the topic or genre effect  
(Vajjala & Meurers 2014b)

- We used our feature set to train a model that identifies age-specific TV programs.
- Data: subtitles from 4 BBC channels, that are annotated with three age-groups.
- Result: 96% classification accuracy
- Analysis:
  - single most predictive feature: average AoA of words
  - But accuracy is not reduced if this feature is removed.
  → The classifier is informed by a range of characteristics, not just a single, dominating one.
Does this approach work on real-world texts? (Vajjala & Meurers 2014c)

- How well does our readability model generalize?
- Do these features generalize for other datasets?
- Is there a topic or genre effect?
- What happens if we analyze sentences?
Documents to sentences?
Readability at sentence level

- Does a readability model identify differences between unsimplified and simplified versions of sentences?
  - Yes, e.g., consider (sentences from OneStopEnglish.com):

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Actual reading level</th>
<th>Reading level assigned by our model</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Beijing, mourners and admirers made their way to lay flowers and light candles at the Apple Store.</td>
<td>Advanced</td>
<td>3.6</td>
</tr>
<tr>
<td>In Beijing, mourners and admirers came to lay flowers and light candles at the Apple Store.</td>
<td>Intermediate</td>
<td>2.3</td>
</tr>
<tr>
<td>In Beijing, people went to the Apple Store with flowers and candles.</td>
<td>Elementary</td>
<td>1.0</td>
</tr>
</tbody>
</table>

- Why is such sentence level annotation needed?
  - for us: performing and evaluating text simplification
Readability at sentence level
How do we go about this?

► Can our model detect readability at sentence level? → Yes, because we do not have any discourse features.
► How do we do that?
  ► Approach 1: train a model on sentence level data.
  ► Approach 2: use document level model on sentences.
► Approach 1 did not work well. So we explored approach 2.
  ► We first studied the distribution of reading levels between sentence pairs.
  ► Later used the scores of a document level model to train a sentence ranking model (work in progress).
Readability at sentence level

Two Datasets

- A three level (advanced/intermediate/elementary) sentence aligned corpus.
  - compiled by us from OneStopEnglish.com, consisting of simplified versions of articles from The Guardian
Readability at sentence level
What did we learn? (Vajjala & Meurers 2014a)

- Simplification is relative
  - A simplified sentence is simpler than its unsimplified version, but can be harder than another sentence.

- Our approach is particularly successful in classifying the relative reading level of harder sentences.

- It promises to be useful for identifying particularly relevant targets for simplification
  - and to evaluate simplifications given specific readability constraints.
Our Reading level predictor in action

Welcome to our readability assessment tool

You can use this interface to determine the reading level of a text by copying the text itself to the text area, uploading the text file or by supplying the URL to the text. Please keep in mind that at the moment we only support texts written in English.

Text/Link  File

Please enter your Text

http://en.wikipedia.org/wiki/Louvain-la-Neuve

Submit  Analysis as txt

Analysis Results
- Total Score
- 14-16 Years
- Sentence Length
- 11-13 Years
- Number ofClauses
- 8-9 Years
- Morph. Complex Words
- 8-9 Years
- Average Age of Acquisition
- 7-8 Years
- Word Imageability
- 8-9 Years

Explanation
The given input is suitable for readers of ages: 14-16

The web-interface is being developed by Ido Friman and will be available soon.
A search engine filtering by vocabulary difficulty

An extension of the Language-Aware Search Engine (??).
Readability Assessment beyond English

- Will this approach work for another language?
- Yes, we experimented with German (Hancke et al. 2012).
We created a two class German corpus from two websites containing articles on similar topics:

- Website for children: GEOlino (http://www.geolino.de)
- Website for adults: GEO (http://www.geo.de)
- They are educational monthly magazines covering articles related to technology, nature etc.
- GEOlino not a simplified version of GEO. 
  - Content is created separately for child readers.

Modeling: Binary classification.
Readability Assessment for German
Features and Model evaluation (Hancke, Meurers & Vajjala 2012)

- **Features:**
  - **General:** lexical, syntactic, language model features
  - **Language specific:** based on inflectional and derivational morphology of German

- **Model performance:**
  - Binary classification: 90% accuracy
  - Morphological features alone perform at 85% accuracy.
Current and Future Work

- We are working on combining regression and ranking for determining readability at sentence level,
  - and using this information for automatic text simplification.

- Our related projects in LEAD:
  - Analyzing German school text books across grade levels and school types using readability features
    - with Karin Berendes, Detmar Meurers, Doreen Bryant
  - Studying the cognitive correlates of readability using eye-tracking experiments.
    - with Detmar Meurers, Katharina Scheiter, Alexander Eitel
  - Towards Appropriate Reading Material for Bilingual Classrooms: Evaluating the role of linguistic complexity analysis and text simplification in authentic contexts
    - with Detmar Meurers, Kathrin Jonkmann, Jörg-U. Keßler
Thank you!

- Questions? :-)
- Contact: sowmya@sfs.uni-tuebingen.de
- Bibliography: http://purl.org/net/readability-bib.html
References


