Real Time Web Text Classification and Analysis of Reading Difficulty

Eleni Miltsakaki & Audrey Troutt
University of Pennsylvania
Summary

- Two system components to support struggling readers
  - **Read-X**: searches for text on the web, classifies it thematically and analyzes its reading difficulty
  - **TOREADOR**: takes text as input, highlights words predicted to be difficult given the reader’s prior familiarity with thematic area
Outline

- Motivation and design principles
- Description of Read-X
  - Web search
  - Text extraction
  - Text classification
  - Readability formulas
- Modeling the reader
  - Prior thematic familiarity
  - Description of TOREADOR
- Related work
- Future work
Outline

- Motivation and design principles
- Description of Read-X
  - Web search
  - Text extraction
  - Text classification
  - Readability formulas
- Modeling the reader
  - Description of TOREADOR
- Related work
- Future work
Struggling readers

- 29% of high school seniors below basic achievement in reading in 2005
  
  *(U.S. Department of Education 2005)*

- Lack of age and interest appropriate reading material for adolescent and low level readers.

- Look at the web!
Design principles

- Easy access and use
  - Run from the web, no installation, no manual, no fees

- Text retrieval and analysis in real time
  - Accessing the web directly
  - No need for updating pre-processed database

- Empowering the educator
  - by providing a tool that s/he can use to build his/her own curriculum
Outline

- Motivation
- Applications
- **Read-X**
  - Web search
  - Text extraction
  - Readability analysis
  - Text classification
- Modeling the reader
  - Description of TOREADOR
- Related work
- Future work
Read-X

- Searches the web on keyword prompt

- Extracts the human-readable text from the html, xml, doc or PDF document stored at each URL.

- Returns text classified:
  - Thematically
  - Reading level
Text classification

- **Corpus**
  - 3.4 m words, mostly from NetTrekker, manually tagged

- Experimented with three classifiers
  - Naïve Bayes
  - MaxEnt
  - MIRA (*Crammer et al* 2008)

- Three levels of granularity
  - 3-way (3 themes)
  - 8-way (8 themes)
  - 41-way (41 themes)

*NB and MaxEnt using MALLET’s tools ([http://mallet.cs.umass.edu](http://mallet.cs.umass.edu))*
## Results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>3-way</th>
<th>8-way</th>
<th>41-way</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>88.23%</td>
<td>66%</td>
<td>30%</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>92.86%</td>
<td>78%</td>
<td>66%</td>
</tr>
<tr>
<td>MIRA</td>
<td>N/A</td>
<td>76%</td>
<td>58%</td>
</tr>
</tbody>
</table>
Readability analysis

- This version of Read-X computes traditional readability formulas

- Lix, Rix, Coleman-Liau
  
  *(Anderson 1983, Coleman & Liau 1975)*
Outline

- Motivation
- Applications
- Read-X
  - Web search
  - Text extraction
  - Readability analysis
  - Text classification
- Modeling the reader
  - Description of TOREADOR
- Related work
- Future work
Modeling the reader

- Some reader variables
  - Familiarity with thematic area
  - Reading ability
  - Interest/Motivation
Modeling familiarity

- *Word frequencies per thematic area*

- Typical word frequency indices used by test developers and educators are computed from corpora with mixed text.
<table>
<thead>
<tr>
<th>Word</th>
<th>Arts Freq</th>
<th>Career and Business Freq</th>
<th>Literature Freq</th>
<th>Philosophy Freq</th>
<th>Science Freq</th>
<th>Social Studies Freq</th>
<th>SportHealth Freq</th>
<th>Technology Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>musical</td>
<td>166</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seem</td>
<td>1398</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trees</td>
<td>174</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>behavior</td>
<td>258</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>players</td>
<td>508</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>software</td>
<td>584</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bacteria</td>
<td>641</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>states</td>
<td>247</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>psychoanalytic</td>
<td>222</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>player</td>
<td>435</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>income</td>
<td>1201</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>myself</td>
<td>158</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>knowledge</td>
<td>158</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>league</td>
<td>443</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>computers</td>
<td>432</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>computer</td>
<td>432</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>software</td>
<td>584</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>site</td>
<td>333</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>video</td>
<td>308</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>player</td>
<td>435</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>games</td>
<td>303</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>football</td>
<td>359</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>games</td>
<td>320</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>used</td>
<td>220</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>resources</td>
<td>171</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>caesar</td>
<td>1059</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>truth</td>
<td>130</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>years</td>
<td>472</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>psychology</td>
<td>157</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>games</td>
<td>320</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>used</td>
<td>220</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>systems</td>
<td>200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>programming</td>
<td>174</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>artist</td>
<td>104</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>asked</td>
<td>153</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>existence</td>
<td>115</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>species</td>
<td>341</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>society</td>
<td>131</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>years</td>
<td>263</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seasons</td>
<td>224</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>using</td>
<td>172</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>engineerin</td>
<td>g</td>
<td>170</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TOREADOR

• TOREADOR highlights words expected to be difficult for the reader.

• Predicts difficult words per grade level and thematic familiarity
TOREADOR

To read or not to read—that is the question!

Enter Text Here  Read Text Here

With Mentors at Their Sides, Girls in Need Write Their Stories and Find New Lives
By J. COURTNEY SULLIVAN

On Saturdays during the school year and all week in the summer, PinChang Huang, 16, leaves her home in Queens just after dawn and boards a crowded van bound for a nail salon on Long Island.

Through a long workday, she gives manicures, pedicures and massages, and observes her clients at the Aroma Spa with a careful eye. Older women are most prone to yell if you make a mistake. Customers who read books tend to tip the best.

PinChang has not seen her mother since she came to New York with her father and brother four years ago from a small village in China. She spoke no English and had no friends, and all the buildings looked the same to her, so she often walked into the wrong apartment complex on her way home from school. To ward off frustration and loneliness, she started keeping a journal.

"I wrote down everything I saw, everything that made me happy or upset," she said. "I wrote the things I wished I could say out loud." In front of a packed

Reader's grade level 02 03 04 05 06 07 08 09 10 11 12 13+
Familiar subjects to reader: Literature Sports Science

Noun

- S (a) manicure (professional care for the hands and fingernails)

Verb

- S (v) manicure (trim carefully and nearly) "manicure fingernails"
- S (v) manicure (care for (one's hand) by cutting and shaping the nails, etc.)
TOREADOR

To read or not to read—that is the question!

Enter Text Here | Read Text Here

With Mentors at Their Sides, Girls in Need Write Their Stories and Find New Lives

By J. COURTNEY SULLIVAN

On Saturdays during the school year and all week in the summer, PinChang Huang, 16, leaves her home in Queens just after dawn and boards a crowded van bound for a nail salon on Long Island. Through a long workday, she gives manicures, pedicures, and massages, and observes her clients at the Aroma Spa with a careful eye. Older women are most prone to yell if you make a mistake. Customers who read books tend to tip the best. PinChang has not seen her mother since she came to New York with her father and brother four years ago from a small village in China. She spoke no English and had no friends, and all the buildings looked the same to her, so she often walked into the wrong apartment complex on her way home from school. To ward off frustration and loneliness, she started keeping a journal. "I wrote down everything I saw, everything that made me happy or upset," she said. "I wrote the things I wished I could say out loud." In front of a packed auditorium at the New School in Manhattan one night recently, she got her chance. PinChang and her mentor, Deborah Kolben, a former managing editor at The Village Voice, read an essay they wrote together about PinChang's getting her first manicure. PinChang spoke about the peculiar sensation of being on the receiving end of an exchange that often makes her feel "like a slave." The reading was hosted by a nonprofit group that pairs high school girls from disadvantaged backgrounds who want to be writers with women who are authors, journalists, playwrights, poets and editors. The group produces an anthology of student writing each spring, and puts on several public readings. The readings are often the first chance girls get to read their own words in front of an audience. "I was so nervous when I stepped onstage," recalled PinChang, a junior at Flushing International High School. "I was shaking. But now I feel like I can say or do anything." Maya Nussbaum, 31, helped found the group, Girls Write Now, 10 years ago when she was a senior majoring in creative writing at Sarah Lawrence College. She knows firsthand how difficult it can be to find a voice in a world that keeps telling you to be quiet.

WordNet home page
Related systems

- **NetTrekker** *(commercial product)*
  - manually built database
  - organized by readability, grade level and subject area

- **REAP Tutor** *(Heilman et al 2006/8)*
  - Intelligent tutoring system
  - Off-line pre-processed database
  - Includes exercises to support ESL vocabulary learning
Current and future work

- Evaluation studies
  - Reading eye-tracking to evaluate predictions of difficult vocabulary
  - Self-paced reading + comprehension question for overall reading difficulty
Current and future work

- Deeper understanding of reading difficulty
  - Syntactic/semantic complexity
    - Syntactic ‘signature’ of different levels of complexity
    - Syntactic proxies for propositional density
  - Perceived coherence
    - Number of introduced topics and length of elaboration
    - Rhetorical structure and intensity of inferencing
Thank you!