Building Readability Lexicons with Unannotated Corpora

Introduction

Our Goals
• Increase coverage and granularity of an existing lexicon for word difficulty
• Use lexicon to provide automatic support to learners

Related work
• Standard readability metrics (Kincaid et al., 1975; Gunning, 1952)
• Text readability classification with lexical features (Collins-Thompson and Callan, 2005; Heilman et al., 2007)
• Deriving readability of lexical items (Kidwell et al. 2009; Li and Feng 2011)
• Creation and evaluation of other kinds of lexicons (Turney and Littman, 2003; Brooke et al., 2010; Taboada et al., 2011)

Method

Basic Procedure
• Extract relevant features for each word
• Linear combination of features to get a measure of difficulty

Simple Features
• From standard readability metrics
• Includes:
  • Term frequency (log) in corpus
  • Word length
  • Syllable length

Document Features
• Calculated at the document level, averaged across documents
• For example, the average word length is average length of words in documents \( D_i \) that a given word appears in:

\[
AWL(w) = \frac{1}{|D_i|} \sum_{d \in D_i} \frac{|w|}{|d|}
\]
• Includes:
  • Avg. word length
  • Avg. sentence length
  • Avg. type-token ratio
  • Avg. lexical density

Co-occurrence Features
• Apply latent semantic analysis (Landauer and Dumais, 1997)
• Value of feature is (normalized average cosine distance of word vectors \( w \) to positive \( P \) and negative \( N \) seed terms):

\[
\text{AVC}(w) = \frac{1}{|P|+|N|} \sum_{p \in P} \cos(w, p) + \sum_{n \in N} \cos(w, n)
\]
• Includes:
  • Formality seed words (Brooke et al., 2010)
  • Childish/abstract seed words
  • Seeds from Difficulty lexicon

Linear combination
• Co-efficients selected using machine learning (Witten and Frank, 2005)
• Linear regression
• For training, beginner words 0.0, intermediate 0.5, advanced 1.0
• Linear SVM
• Use relative rather than absolute judgments
• Other algorithms

Resources

Difficulty Lexicon
• 15,308 words from other lists (e.g. Dolch, 1948) and age-graded corpora
• Manually assigned to 3 difficulty levels:
  • Beginner (e.g. boat, arrow, lizard, earn, afternoon)
  • Intermediate (e.g. motto, survey, intestine, conflict)
  • Advanced (e.g. contingency, scoff, illegitimate, myriad)
• Filtered, 500 testing and 300 training/development per level
• Each word paired with another word from each level to create 4500/2700 pairs

Crowdflower Annotation
• For each pair, ask workers which word was learned first (first, second, or same)
• 5 judgments, majority used, or same if conflict
• Quality control
• Publicly available blog corpus, the ICWSM 2009 (Burton et al., 2009)
• 1.3 billion tokens, mixed register

Evaluation

Evaluation of Annotations
• Moderate agreement among Crowdflower workers (56.6%)
• High (72.5%) for extreme categories, low (46%) for same categories
• 63.1% agreement between Crowdflower and Difficulty lexicon
• Some judgment relatively rare in Crowdflower
• If some judgments are disregarded, agreement is high (91.0%)
• Our current lexicon lacks fine-grainedness

Evaluation of Automatic Lexicon
• Only use non-same judgements
• Crowdflower more difficult
• More subtle distinctions
• Frequency important for Crowdflower
• Few individual features are poor
• But: syllable, type-token
• Co-occurrence features redundant
• With each other
• With Document features
• Otherwise, major boost from combining
• Linear regression and SVM similar
• SVM only needs relative annotation
• 91.2% for pairs where both agreed

Discussion
• High granularity, low reliability?
• Co-occurrence advantages
• Capturing child/adult vocabulary difference
• E.g. dollhouse/emergence
• Word length not for all languages
• Potentially useful for L2 learners

Conclusion
• Blog texts help with expansion of our lexicon of difficulty
• Useful features go beyond term frequency

References and Acknowledgments

Collins-Thompson, Kevyn, and Michael J. Thoms. 2009. Statistical estimation of word occurrence features redundant