Julian Brooke* Vivian Tsang† David Jacob† Fraser Shein*† Graeme Hirst *

*Department of Computer Science, University of Toronto: {jbrooke,gh}@cs.toronto.edu †Quillsoft Ltd: {vtsang,djacob,fshein}@quillsoft.com

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_w) that a given word appears in:

$$AWL(w) = |D_w|^{-1} \sum_{d \in D_w} \frac{\sum_{i=0}^{|d|} length(d_i)}{|d|}$$

- 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 ($\mathbf{w}$) to positive (P) and negative (N) seed terms:

$$V(\mathbf{w}) = \frac{\sum_{\mathbf{p} \in P} \cos(\theta(\mathbf{w}, \mathbf{p}))}{|P|} - \frac{\sum_{\mathbf{n} \in N} \cos(\theta(\mathbf{w}, \mathbf{n}))}{|N|}$$

- 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. *coat*, *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

Corpus

- 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
- *Same* judgment relatively rare in Crowdflower
  - If *same* 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

Agreement (%) of automated methods with manual resources on pairwise comparison task (Diff. = Difficulty lexicon, CF = Crowdflower)

Features	Resource	
	Diff.	CF
Simple		
Syllable Length	62.5	54.9
Word Length	68.8	62.4
Term Frequency	69.2	70.7
Document		
Avg. Word Length	74.5	66.8
Avg. Sentence Length	73.5	65.9
Avg. Type-Token Ratio	47.0	50.0
Avg. Lexical Density	56.1	54.7
Co-occurrence Features		
Formality	74.7	66.5
Childish	74.2	65.5
Difficulty	75.7	66.1
Linear Combinations		
Simple	79.3	75.0
Document	80.1	70.8
Co-occurrence	76.0	67.0
Document+Co-occurrence	80.4	70.2
Simple+Document	87.4	79.1
Simple+Co-occurrence	86.7	78.2
All	87.6	79.5
All (SVM)	87.1	79.2

Discussion

- High granularity, low reliability?
- Co-occurrence advantages
  - Capturing child/adult vocab difference
    - *E.g. dollhouse/emergence*
  - Word length not for all languages
  - Potentially useful for L2 learner needs

Conclusion

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

References and Acknowledgments

Brooke, Julian, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*.

Burton, Kevin, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. *In Proceedings of the Third Annual Conference on Weblogs and Social Media* (ICWSM '09)

Collins-Thompson, Kevyn and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science Technology*, 56(13):1448–1462.

Dolch, Edward William. 1948. *Problems in Reading*. The Garrard Press.

Gunning, Robert. 1952. *The Technique of Clear Writing*. McGraw-Hill.

Heilman, Michael J., Kevyn Collins-Thompson, and Jamie Callan. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of the Conference of the North American Chapter of Association for Computational Linguistics* (NAACL-HLT '07).

Kidwell, Paul, Guy Lebanon, and Kevyn Collins-Thompson. 2009. Statistical estimation of word acquisition with application to readability prediction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (EMNLP'09), 900–909.

Kincaid, J. Peter, Robert. P. Fishburne Jr., Richard L. Rogers, and Brad. S. Chissom. 1975. Derivation of new readability formulas for Navy enlisted personnel. Research Branch Report 8-75, Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

Landauer, Thomas K. and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240

Li, Hanhong and Alex C. Feng. 2011. Age tagging and word frequency for learners' dictionaries. In Harald Baayan, John Newman and Sally Rice, editors, *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. Rodopi.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manifred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Turney, Peter and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346.

Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

