Building Readability Lexicons with Unannotated Corpora

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

[*]Department of Computer Science
University of Toronto
{jbrooke,gh}@cs.toronto.edu

[†]Quillsoft Ltd.
Toronto, Canada
{vtsang, djacob, fshein}@quillsoft.ca

Abstract

Lexicons of word difficulty are useful for various educational applications, including readability classification and text simplification. In this work, we explore automatic creation of these lexicons using methods which go beyond simple term frequency, but without relying on age-graded texts. In particular, we derive information for each word type from the readability of the web documents they appear in and the words they co-occur with, linearly combining these various features. We show the efficacy of this approach by comparing our lexicon with an existing coarse-grained, low-coverage resource and a new crowdsourced annotation.

1 Introduction

With its goal of identifying documents appropriate to readers of various proficiencies, automatic analysis of readability is typically approached as a text-level classification task. Although at least one popular readability metric (Dale and Chall, 1995) and a number of machine learning approaches to readability rely on lexical features (Si and Callan, 2001; Collins-Thompson and Callan, 2005; Heilman et al., 2007; Petersen and Ostendorf, 2009; Tanaka-Ishii et al., 2010), the readability of individual lexical items is not addressed directly in these approaches. Nevertheless, information about the difficulty of individual lexical items, in addition to being useful for text readability classification (Kidwell et al., 2009), can be applied to other tasks, for instance lexical simplification (Carroll et al., 1999; Burstein et al., 2007).

Our interest is in providing students with educational software that is sensitive to the difficulty of particular English expressions, providing proactive support for those which are likely to be outside a reader's vocabulary. However, our existing lexical resource is coarse-grained and lacks coverage. In this paper, we explore the extent to which an automatic approach could be used to fill in the gaps of our lexicon. Prior approaches have generally depended on some kind of age-graded corpus (Kidwell et al., 2009; Li and Feng, 2011), but this kind of resource is unlikely to provide the coverage that we require; instead, our methods here are based on statistics from a huge web corpus. We show that frequency, an obvious proxy for difficulty, is only the first step; in fact we can derive key information from the documents that words appear in and the words that they appear with, information that can be combined to give high performance in identifying relative difficulty. We compare our automated lexicon against our existing resource as well as a crowdsourced annotation.

2 Related Work

Simple metrics form the basis of much readability work: most involve linear combinations of word length, syllable count, and sentence length (Kincaid et al., 1975; Gunning, 1952), though the popular Dale-Chall reading score (Dale and Chall, 1995) is based on a list of 3000 'easy' words; a recent review suggests these metrics are fairly interchangeable (van Oosten et al., 2010). In machine-learning classification of texts by grade level, unigrams have been found to be reasonably effective for this task, outperforming readability metrics (Si and Callan, 2001; Collins-Thompson and Callan, 2005). Var-

NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for target reader populations (PITR 2012)., pages 33–39,
Montréal, Canada, June 7, 2012. ©2012 Association for Computational Linguistics

ious other features have been explored, including parse (Petersen and Ostendorf, 2009) and coherence features (Feng et al., 2009), but the consensus seems to be that lexical features are the most consistently useful for automatic readability classification, even when considering non-native readers (Heilman et al., 2007).

In the field of readability, the work of Kidwell et al. (2009) is perhaps closest to ours. Like the above, their goal is text readability classification, but they proceed by first deriving an age of acquisition for each word based on its statistical distribution in age-annotated texts. Also similar is the work of Li and Feng (2011), who are critical of raw frequency as an indicator and instead identify core vocabulary based on the common use of words across different age groups. With respect to our goal of lowering reliance on fine-grained annotation, the work of Tanaka-Ishii et al. (2010) is also relevant; they create a readability system that requires only two general classes of text (easy and difficult), other texts are ranked relative to these two classes using regression.

Other lexical acquisition work has also informed our approach here. For instance, our co-occurrence method is an adaption of a technique applied in sentiment analysis (Turney and Littman, 2003), which has recently been shown to work for formality (Brooke et al., 2010), a dimension of stylistic variation that seems closely related to readability. Taboada et al. (2011) validate their sentiment lexicon using crowdsourced judgments of the relative polarity of pairs of words, and in fact crowd sourcing has been applied directly to the creation of emotion lexicons (Mohammad and Turney, 2010).

3 Resources

Our primary resource is an existing lexicon, previously built under the supervision of the one of authors. This resource, which we will refer to as the Difficulty lexicon, consists of 15,308 words and expressions classified into three difficulty categories: beginner, intermediate, and advanced. Beginner, which was intended to capture the vocabulary of early elementary school, is an amalgamation of various smaller sources, including the Dolch list (Dolch, 1948). The intermediate words, which include words learned in late elementary and middle school, were extracted from Internet-published texts written by students at these grade levels, and then filtered manually. The advanced words began as a list of common words that were in neither of the original two lists, but they have also been manually filtered; they are intended to reflect the vocabulary understood by the average high school student. Table 1 contains some examples from each list.

Table 1: Examples from the Difficulty lexicon

Beginner
coat, away, arrow, lizard, afternoon, rainy, carpet, earn, hear, chill
Intermediate
bale, campground, motto, intestine, survey, regularly, research, conflict
Advanced
contingency, scoff, characteristic, potent, myriad, detracted, illegitimate, overture

For our purposes here, we only use a subset of the Difficulty lexicon: we filtered out inflected forms, proper nouns, and words with non-alphabetic components (including multiword expressions) and then randomly selected 500 words from each level for our test set and 300 different words for our development/training set. Rather than trying to duplicate our arbitrary three-way distinction by manual or crowdsourced means, we instead focused on the relative difficulty of individual words: for each word in each of the two sets, we randomly selected three comparison words, one from each of the difficulty levels, forming a set of 4500 test pairs (2700 for the development set): $1/3$ of these pairs are words from the same difficulty level, $4/9$ are from adjacent difficulty levels, and the remaining $2/9$ are at opposite ends of our difficulty spectrum.

Our crowdsourced annotation was obtained using Crowdflower, which is an interface built on top of Mechanical Turk. For each word pair to be compared, we elicited 5 judgments from workers. Rather than frame the question in terms of difficulty or readability, which we felt was too subjective, we instead asked which of the two words the worker thought he or she learned first: the worker could choose either word, or answer “about the same time”. They were instructed to choose the word they did know if one of the two words was unknown, and "same" if both were unknown. For our evaluation, we took the majority judgment as the gold standard; when there was no majority judgment, then the words were considered "the same". To increase the likelihood that our workers were native speakers of English, we required that the responses come from the US or Canada. Before running our main set, we ran several smaller test runs and manually inspected them for quality; although there were outliers, the majority of the judgments seemed reasonable.

Our corpus is the ICWSM Spinn3r 2009 dataset (Burton et al., 2009). We chose this corpus because it was used by Brooke et al. (2010) to derive a lexicon of formality; they found that it was more effective for these purposes than smaller mixed-register corpora like the BNC. The ICWSM 2009, collected over several weeks in 2008, contains about 7.5 million blogs, or 1.3 billion tokens, including well over a million word types (more than 200,000 of which which appear at least 10 times). We use only the documents which have at least 100 tokens. The corpus has been tagged using the TreeTagger (Schmid, 1995).

4 Automatic Lexicon Creation

Our method for lexicon creation involves first extracting a set of relevant numerical features for each word type. We can consider each feature as defining a lexicon on its own, which can be evaluated using our test set. Our features can be roughly broken into three types: simple features, document readability features, and co-occurrence features. The first of these types does not require much explanation: it includes the length of the word, measured in terms of letters and syllables (the latter is derived using a simple but reasonably accurate vowel-consonant heuristic), and the log frequency count in our corpus.[1]

The second feature type involves calculating simple readability metrics for each document in our corpus, and then defining the relevant feature for the word type as the average value of the metric for all the documents that the word appears in. For example, if D_w is the set of documents where word type w appears and d_i is the ith word in a document d, then the *document word length* (DWL) for w can be defined as follows:

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

Other features calculated in this way include: the document sentence length, that is the average token length of sentences; the document type-token ratio[2]; and the document lexical density, the ratio of content words (nouns, verbs, adjectives, and adverbs) to all words.

The co-occurence features are inspired by the semi-supervised polarity lexicon creation method of Turney and Littman (2003). The first step is to build a matrix consisting of each word type and the documents it appears in; here, we use a binary representation, since the frequency with which a word appears in a particular document does not seem directly relevant to readability. We also do not remove traditional stopwords, since we believe that the use of certain common function words can in fact be good indicators of text readability. Once the matrix is built, we apply latent semantic analysis (Landauer and Dumais, 1997); we omit the mathematical details here, but the result is a dimensionality reduction such that each word is represented as a vector of some k dimensions. Next, we select two sets of seed words (P and N) which will represent the ends of the spectrum which we are interested in deriving. We derive a feature value V for each word by summing the cosine similarity of the word vector with all the seeds:

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

We further normalize this to a range of 1 to -1, centered around the core vocabulary word *and*. Here, we try three possible versions of P and N: the first, Formality, is the set of words used by Brooke et al. (2010) in their study of formality, that is, a

[1] Though it is irrelevant when evaluating the feature alone, the log frequency was noticeably better when combining frequency with other features.

[2] We calculate this using only the first 100 words of the document, to avoid the well-documented influence of length on TTR.

set of slang and other markers of oral communication as *N*, and a set of formal discourse markers and adverbs as *P*, with about 100 of each. The second, Childish, is a set of 10 common 'childish' concrete words (e.g. *mommy*, *puppy*) as *N*, and a set of 10 common abstract words (e.g. *concept*, *philosophy*) as *P*. The third, Difficulty, consists of the 300 beginner words from our development set as *N*, and the 300 advanced words from our development set as *P*. We tested several values of k for each of the seed sets (from 20 to 500); there was only small variation so here we just present our best results for each set as determined by testing in the development set.

Our final lexicon is created by taking a linear combination of the various features. We can find an appropriate weighting of each term by taking them from a model built using our development set. We test two versions of this: by default, we use a linear regression model where for training beginner words are tagged as 0, advanced words as 1, and intermediate words as 0.5. The second model is a binary SVM classifier; the features of the model are the difference between the respective features for each of the two words, and the classifier predicts whether the first or second word is more difficult. Both models were built using WEKA (Witten and Frank, 2005), with default settings except for feature normalization, which must be disabled in the SVM to get useful weights for the linear combination which creates our lexicon. In practice, we would further normalize our lexicon; here, however, this normalization is not relevant since our evaluation is based entirely on relative judgments. We also tested a range of other machine learning algorithms available in WEKA (e.g. decision trees and MaxEnt) but the crossvalidated accuracy was similar to or slightly lower than using a linear classifier.

5 Evaluation

All results are based on comparing the relative difficulty judgments made for the word pairs in our test set (or, more often, some subset) by the various sources. Since even the existing Difficulty lexicon is not entirely reliable, we report agreement rather than accuracy. Except for agreement of Crowdflower workers, agreement is the percentage of pairs where the sources agreed as compared to the total number of pairs. For agreement between Crowdflower workers, we follow Taboada et al. (2011) in calculating agreement across all possible pairings of each worker for each pair. Although we considered using a more complex metric such as Kappa, we believe that simple pairwise agreement is in fact equally interpretable when the main interest is relative agreement of various methods; besides, Kappa is intended for use with individual annotators with particular biases, an assumption which does not hold here.

To evaluate the reliability of our human-annotated resources, we look first at the agreement within the Crowdflower data, and between the Crowdflower and our Difficulty lexicon, with particular attention to within-class judgments. We then compare the predictions of various automatically extracted features and feature combinations with these human judgments; since most of these involve a continuous scale, we focus only on words which were judged to be different.[3] For the Difficulty lexicon (Diff.), the n in this comparison is 3000, while for the Crowdflower (CF) judgments it is 4002.

6 Results

We expect a certain amount of noise using crowdsourced data, and indeed agreement among Crowdflower workers was not extremely high, only 56.6% for a three-way choice; note, however, that in these circumstances a single worker disagreeing with the rest will drop pairwise agreement in that judgement to 60%.[4] Tellingly, average agreement was relatively high (72.5%) for words on the extremes of our difficulty spectrum, and low for words in the same difficulty category (46.0%), which is what we would expect. As noted by Taboada et al. (2011), when faced with a pairwise comparison task, workers tend to avoid the "same" option; instead, the proximity of the words on the underlying spectrum is reflected in disagreement. When we compare the crowdsourced judgements directly to the Difficulty lexicon, base

[3] A continuous scale will nearly always predict some difference between two words. An obvious approach would be to set a threshold within which two words will be judged the same, but the specific values depend greatly on the scale and for simplicity we do not address this problem here.

[4] In 87.3% of cases, at least 3 workers agreed; in 56.2% of cases, 4 workers agreed, and in 23.1% of cases all 5 workers agreed.

agreement is 63.1%. This is much higher than chance, but lower than we would like, considering these are two human-annotated sources. However, it is clear that much of this disagreement is due to "same" judgments, which are three times more common in the Difficulty lexicon-based judgments than in the Crowdflower judgments (even when disagreement is interpreted as a "same" judgment). Pairwise agreement of non-"same" judgments for word pairs which are in the same category in the Difficultly lexicon is high enough (45.9%)[5] for us to conclude that this is not random variation, strongly suggesting that there are important distinctions within our difficulty categories, i.e. that it is not sufficiently fine-grained. If we disregard all words that are judged as same in one (or both) of the two sources, the agreement of the resulting word pairs is 91.0%, which is reasonably high.

Table 2 contains the agreement when feature values or a linear combination of feature values are used to predict the readability of the unequal pairs from the two manual sources. First, we notice that the Crowdflower set is obviously more difficult, probably because it contains more pairs with fairly subtle (though noticeable) distinctions. Other clear differences between the annotations: whereas for Crowdflower frequency is the key indicator, this is not true for our original annotation, which prefers the more complex features we have introduced here. A few features did poorly in general: syllable count appears too coarse-grained to be useful on its own, lexical density is only just better than chance, and type-token ratio performs at or below chance. Otherwise, many of the features within our major types give roughly the same performance individually.

When we combine features, we find that simple and document features combine to positive effect, but the co-occurrence features are redundant with each other and, for the most part, the document features. A major boost comes, however, from combining either document or co-occurrence features with the simple features; this is especially true for our Difficulty lexicon annotation, where the gain is 7% to 8 percentage points. It does not seem to matter very much whether the weights of each feature are determined by pairwise classifier or by linear regression: this is interesting because it means we can train a model to create a readability spectrum with only pairwise judgments. Finally, we took all the 2500 instances where our two annotations agreed that one word was more difficult, and tested our best model against only those pairs. Results using this selective test set were, unsurprisingly, higher than those of either of the annotations alone: 91.2%, which is roughly the same as the original agreement between the two manual annotations.

Table 2: 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		
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	87.5	79.1
Simple+Co-occurrence	86.7	78.2
All	**87.6**	**79.5**
All (SVM)	87.1	79.2

7 Discussion

Word difficulty is a vague concept, and we have admittedly sidestepped a proper definition here: instead, we hope to establish a measure of reliability in judgments of 'lexical readability' by looking for agreement across diverse sources of information. Our comparison of our existing resources with

[5] Random agreement here is 33.3%.

crowdsourced judgments suggests that some consistency is possible, but that granularity is, as we predicted, a serious concern, one which ultimately undermines our validation to some degree. An automatically derived lexicon, which can be fully continuous or as coarse-grained as needed, seems like an ideal solution, though the much lower performance of the automatic lexicon in predicting the more fine-grained Crowdflower judgments indicates that automatically-derived features are limited in their ability to deal with subtle differences. However, a visual inspection of the spectrum created by the automatic methods suggests that, with a judicious choice of granularity, it should be sufficient for our needs. In future work, we also intend to evaluate its use for readability classification, and perhaps expand it to include multiword expressions and syntactic patterns.

Our results clearly show the benefit of combining multiple sources of information to build a model of word difficulty. Word frequency and word length are of course relevant, and the utility of the document context features is not surprising, since they are merely a novel extension of existing proxies for readability. The co-occurrence features were also useful, though they seem fairly redundant and slightly inferior to document features; we posit that these features, in addition to capturing notions of register such as formality, may also offer semantic distinctions relevant to the acquisition process. For instance, children may have a large vocabulary in very concrete domains such as animals, including words (e.g. *lizard*) that are not particularly frequent in adult corpora, while very common words in other domains (such as the legal domain) are completely outside the range of their experience. If we look at some of the examples which term frequency alone does not predict, they seem to be very much of this sort: *dollhouse*/*emergence*, *skirt*/*industry*, *magic*/*system*. Unsupervised techniques for identifying semantic variation, such as LSA, can capture these sorts of distinctions. However, our results indicate that simply looking at the readability of the texts that these sort of words appear in (i.e. our document features) is mostly sufficient, and less than 10% of the pairs which are correctly ordered by these two feature sets are different. In any case, an age-graded corpus is definitely not required.

There are a few other benefits of using word co-occurrence that we would like to touch on, though we leave a full exploration for future work. First, if we consider readability in other languages, each language may have different properties which render proxies such as word length much less useful (e.g. ideographic languages like Chinese or agglutinative languages like Turkish). However, word (or lemma) co-occurrence, like frequency, is essentially a universal feature across languages, and thus can be directly extended to any language. Second, if we consider how we would extend difficulty-lexicon creation to the context of adult second-language learners, it might be enough to adjust our seed terms to reflect the differences in the language exposure of this population, i.e. we would expect difficulty in acquiring colloquialisms that are typically learned in childhood but are not part of the core vocabulary of the adult language.

8 Conclusion

In this paper, we have presented an automatic method for the derivation of a readability lexicon relying only on an unannotated word corpus. Our results show that although term frequency is a key feature, there are other, more complex features which provide competitive results on their own as well as combining with term frequency to improve agreement with manual resources that reflect word difficulty or age of acquisition. By comparing our manual lexicon with a new crowdsourced annotation, we also provide a validation of the resource, while at the same time highlighting a known issue, the lack of fine-grainedness. Our manual lexicon provides a solution for this problem, albeit at the cost of some reliability. Although our immediate interest is not text readability classification, the information derived could be applied fairly directly to this task, and might be particularly useful in the case when annotated texts are not avaliable.

Acknowledgments

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

References

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

Jill Burstein, Jane Shore, John Sabatini, Yong-Won Lee, and Matthew Ventura. 2007. The automated text adaptation tool. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '07), Software Demonstrations*, pages 3–4.

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

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying English text for language impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 269–270.

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

Edgar Dale and Jeanne Chall. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge, MA.

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

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09)*, pages 229–237.

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

Michael J. Heilman, Kevyn Collins, 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)*.

Paul Kidwell, 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)*, pages 900–909.

J. Peter Kincaid, 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.

Thomas K. Landauer 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.

Hanhong Li 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.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23(1):89–106.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop*, pages 47–50.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM '01)*, pages 574–576.

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

Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. 2010. Sorting texts by readability. *Computational Linguistics*, 36(2):203–227.

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

Philip van Oosten, Dries Tanghe, and Veronique Hoste. 2010. Towards an improved methodology for automated readability prediction. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10)*.

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