A New Yardstick and Tool for Personalized Vocabulary Building

Thomas K Landauer **Kirill Kireyev** **Charles Panaccione**

Pearson Education,
Knowledge Technologies

{tom.landauer,kirill.kireyev,charles.panaccione}@pearson.com

Abstract

The goal of this research is to increase the value of each individual student's vocabulary by finding words that the student doesn't know, needs to, and is ready to learn. To help identify such words, a better model of how well any given word is expected to be known was created. This is accomplished by using a semantic language model, LSA, to track how every word changes with the addition of more and more text from an appropriate corpus. We define the "maturity" of a word as the degree to which it has become similar to that after training on the entire corpus.

An individual student's average vocabulary level can then be placed on the word-maturity scale by an adaptive test. Finally, the words that the student did or did not know on the test can be used to predict what other words the same student knows by using multiple maturity models trained on random samples of typical educational readings. This detailed information can be used to generate highly customized vocabulary teaching and testing exercises, such as Cloze tests.

1 Introduction

1.1 Why "Vocabulary First"

There are many arguments for the importance of more effective teaching of vocabulary. Here are some examples:

(1) Baker, Simmons, & Kame'enui (1997) found that children who enter school with limited vocabulary knowledge grow much more discrepant over time from their peers who have rich vocabulary knowledge.

(2.) Anderson & Freebody (1981) found that the number of words in student's meaning vocabularies was the best predictor of how well they comprehend text.

(3) An unpublished 1966 study of the correlation between entering scores of Stanford Students on the SAT found the vocabulary component to be the best predictor of grades in every subject, including science.

(4) The number of words students learn varies greatly, from 0.2 to 8 words per day and from 50 to over 3,000 per year. (Anderson & Freebody,1981)

(5) Printed materials in grades 3 to 9 on average contain almost 90,000, distinct word families and nearly 500,000 word forms (including proper names.) (Nagy & Anderson, 1984).

(6) Nagy and Anderson (1984) found that on average not knowing more than one word in a sentence prevented its tested understanding, and that the probability of learning the meaning of a new word by one encounter on average was less than one in ten.

(7) John B. Carroll's (1993) meta-analysis of factor analyses of measured cognitive ability found the best predictor to be tests of vocabulary.

(8) Hart and Risley's large randomized observational study of the language used in households with young children found that the number of words spoken within hearing of a child was associated with a three-fold difference in vocabulary by school entry.

1.2 The Challenge

Several published sources and inspection of the number of words taught in recent literacy textbooks and online tools suggest that less than 400 words per year are directly tutored in American schools. Thus, the vast majority of vocabulary must be acquired from language exposure, especially from print because the oral vocabulary of daily living is usually estimated to be about 20,000

Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications, pages 27–33,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

words, of which most are known by early school years. But it would obviously be of great value to find a way to make the explicit teaching of vocabulary more effective, and to make it multiply the effects of reading. These are the goals of the new methodologies reported here.

It is also clear that words are not learned in isolation: learning the meaning of a new word requires prior knowledge of many other words, and by most estimates it takes a (widely variable) average of ten encounters in different and separated contexts. (This, by the way, is what is required to match human adult competence in the computational language model used here. Given a text corpus highly similar to that experienced by a language learner, the model learns at very close to the same rate as an average child, and it learns new words as much as four times faster the more old words it knows (Landauer & Dumais, 1997).)

An important aside here concerns a widely circulated inference from the Nagy and Anderson (1984) result that teaching words by presenting them in context doesn't produce enough vocabulary growth to be the answer. The problem is that the experiments actually show only that the inserted target word itself is usually not learned well enough to pass a test. But in the simulations, words are learned a little at a time; exposure to a sentence increases the knowledge of many other words, both ones in the sentence and not. Every encounter with any word in context percolates meaning through the whole current and future vocabulary. Indeed, in the simulator, indirect learning is three to five times as much as direct, and is what accounts for its ability to match human vocabulary growth and passage similarity. Put differently, the helpful thing that happens on encountering an unknown word is not guessing its meaning but its contribution to underlying understanding of language.

However, a vicious negative feedback loop lurks in this process. Learning from reading requires vocabulary knowledge. So the vocabulary-rich get richer and the vocabulary-poor get relatively poorer. Fortunately, however, in absolute terms there is a positive feedback loop: the more words you know, the faster you can learn new ones, generating exponential positive growth. Thus the problem and solution may boil down to increasing the growth parameter for a given student enough to make natural reading do its magic better.

Nonetheless, importantly, it is patently obvious that it matters greatly what words are taught how, when and to which students.

The hypothesis, then, is that a set of tools that could determine what particular words an individual student knows and doesn't, and which ones learned (and sentences understood) would most help other words to be learned by that student might have a large multiplying effect. It is such a toolbox that we are endeavoring to create by using a computational language model with demonstrated ability to simulate human vocabulary growth to a reasonably close approximation. The principal foci are better selection and "personalization" of what is taught and teaching more quickly and with more permanence by application of optimal spacing of tests and practice—into which we will not go here.

1.3 Measuring vocabulary knowledge

Currently there are three main methods for measuring learner vocabulary, all of which are inadequate for the goal. They are:

1. Corpus Frequency. Collect a large sample of words used in the domain of interest, for example a collection of textbooks and readers used in classrooms, text from popular newspapers, a large dictionary or the Internet. Rank the words by frequency of occurrence. Test students on a random subset of, say, the 1,000, 2,000 and 5,000 most frequent words, compute the proportion known at each "level" and interpolate and extrapolate. This is a reasonable method, because frequently encountered words are the ones most frequently needed to be understood.
2. Educational Materials. Sample vocabulary lessons and readings over classrooms at different school grades.
3. Expert Judgments. Obtain informed expert opinions about what words are important to know by what age for what purposes.

Some estimates combine two or more of these approaches, and they vary in psychometric sophistication. For example, one of the most sophisticated, the Lexile Framework, uses Rasch scaling (Rasch, 1980) of a large sample of student vocabulary test scores (probability right on a test, holding student ability constant) to create a difficulty measure for sentences and then infers the difficulty of words, in essence, from the average difficulty of the sentences in which they appear.

The problem addressed in the present project goal is that all of these methods measure only the proportion of tested words known at one or more frequency ranges, in chosen school grades or for particular subsets of vocabulary (e.g. "academic" words), and for a very small subset—those tested - some of the words that the majority of a class knows. What they don't measure is exactly which words in the whole corpus a given student knows and to what extent, or which words would be most important for that student to learn.

A lovely analog of the problem comes from Ernst Rothkopf's (1970) metaphor that everyone passes through highly different "word swarms" each day on their way to their (still highly differentiated) adult literacy.

2 A new metric: Word Maturity

The new metric first applies Latent Semantic Analysis (LSA) to model how representation of individual words changes and grows toward their adult meaning as more and more language is encountered. Once the simulation has been created, an adaptive testing method can be applied to place individual words on separate growth curves - characteristic functions in psychometric terminology. Finally, correlations between growth curves at given levels can be used to estimate the achieved growth of other words.

2.1 How it works in more detail: LSA.

A short review of how LSA works will be useful here because it is often misunderstood and a correct interpretation is important in what follows. LSA models how words combine into meaningful passages, the aspect of verbal meaning we take to be most critical to the role of words in literacy. It does this by assuming that the "meaning" (please bear with the nickname) of a meaningful passage is the sum of the meanings of its words:

```
Meaning of passage =
 {meaning of first wd} +
 {meaning of second word} + …  +
 {meaning of last word}
```

A very large and representative corpus of the language to be modeled is first collected and represented as a term-by-document matrix. A powerful matrix algebra method called Singular Value Decomposition is then used to make every paragraph in the corpus conform to the above objective function—word representations sum to passage representations - up to a best least-squares approximation. A dimensionality-reduction step is performed, resulting in each word and passage meanings represented as a (typically) 300 element real number vector. Note that the property of a vector standing for a word form in this representation is the effect that it has on the vector standing for the passage. (In particular, it is only indirectly a reflection of how similar two words are to each other or how frequently they have occurred in the same passages.) In the result, the vector for a word is the average of the vectors for all the passages in which it occurs, and the vector for a passage is, of course, the average all of its words.

In many previous applications to education, including automatic scoring of essays, the model's similarity to human judgments (e.g. by mutual information measures) has been found to be 80 to 90% as high as that between two expert humans, and, as mentioned earlier, the rate at which it learns the meaning of words as assessed by various standardized and textbook-based tests has been found to closely match that of students. For more details, evaluations and previous educational applications, see (Landauer et al., 2007).

2.2 How it works in more detail: Word Maturity.

Taking LSA to be a sufficiently good approximation of human learning of the meanings conveyed by printed word forms, we can use it to track their gradual acquisition as a function of increasing exposure to text representative in size and content of that which students at successive grade levels read.

Thus, to model the growth of meaning of individual words, a series of sequentially accumulated LSA "semantic spaces" (the collection of vectors for all of the words and passages) are created. Cumulative portions of the corpus thus emulate the growing total amount of text that has been read by a student. At each step, a new LSA semantic space is created from a cumulatively larger subset of the full adult corpus.

Several different ways of choosing the successive sets of passages to be added to the training set have been tried, ranging from ones based on readability metrics (such as Lexiles or DRPs) to en-