Retrieval of Authentic Documents for Reader-Specific Lexical Practice

Jonathan Brown
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213 USA
jonbrown@cs.cmu.edu

Maxine Eskenazi
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213 USA
max@cs.cmu.edu

Abstract

When a teacher gives a reading assignment in today's language learning classrooms, all of the students are almost always reading the same text. Although students have different reading levels, it is impractical for a single teacher to seek out unique texts matched to each student's abilities. In this paper, we describe REAP, a system designed to assign each student individualized readings by combining detailed student and curriculum modelling with the large amount of authentic materials on the Web. REAP is designed to be used as an additional resource in teacher-led classes, as well as to be used by reading comprehension researchers for testing hypotheses on how to improve reading skills for L1 as well as L2 learners. Vocabulary acquisition is the primary factor we use in matching texts to a student's abilities. The system can also prioritise different criteria during the search. For instance, the system can retrieve documents based solely on the vocabulary terms needed to progress toward the next level, thereby focusing on curriculum. REAP can take into account other goals, such as student interests, special topics decided by the teacher, or an upcoming test, all represented as word histograms. This allows teachers to decide what they want the students to focus on each day.

We also describe the contributions of this project, including an open-corpus, authentic-materials approach to reading practice and word-level modelling of norms and student skills. Finally, we describe how learning researchers can use this tool to get fine-grained control over the selection of reading materials, so that they can more easily test a variety of new learning hypotheses.

1 Introduction

This paper describes work on the CMU Language Technologies Institute's REAP project. The goal of REAP is to find appropriate authentic documents for a student learning to read. We present this project in the light of K-12 reading research and refer to how it can be used for L2 (second language) learning.

Although there is increasing realization that it is important to use authentic documents to teach reading, actual practice most often uses prepared text. There are two disadvantages to this. First, the student is not exposed to examples of real language, the language actually used in everyday written communication. Second, the students all get the same text to read. Students who are having trouble with words and students who are well ahead of the others are all reading the same text with little backup for remediation and little chance to leap forward respectively. REAP not only furnishes authentic texts, but also finds texts that are appropriate for the individual reader.

Prior research has established that lexical mastery of core vocabulary at a particular reading level is essential to the development of more complex reading comprehension skills (e.g. Perfetti et al., 1995; Perfetti and Hart, 2001). Traditional approaches to lexical acquisition and fluency are reading practice of texts that stress core vocabulary, and vocabulary tests. They are intended to keep a student in relatively close correspondence with a normative model, in this case the core vocabulary associated with a particular level of reading difficulty. Students with both strong and weak reading abilities acquire their vocabularies in roughly the same order, but differ markedly in their acquisition rates and the overall sizes of their vocabularies (Biemiller and Stonim, 2001).

In the last decade several developments in Computer Science have afforded a different approach to studying lexical acquisition and fluency. One was the successful application of statistical language models to speech recognition tasks and then to a variety of other language tasks. The simplest forms of statistical language models are essentially word histograms that represent the relative frequencies of words in a sample of text; more complex language models represent the frequencies of word sequences of varying lengths ("n-grams"). Very simple language models have been found to be sufficient for creating state-of-the-art information retrieval systems (i.e., "search engines"). One can view a search engine based on a statistical language model as returning text passages that satisfy very specific lexical constraints. When such a search engine is combined with a mammoth source of documents, such as the World Wide Web, the probability of satisfying even relatively specific lexical constraints is reasonably high. This enables a new approach to studying the effects of different strategies for improving lexical acquisition and fluency on reading comprehension. Language models representing norms (e.g., 4th grade reading ability), specific populations (e.g., reading ability of low-income 4th grade students in Pittsburgh), and individuals (e.g., John Doe) can be created automatically from sample texts. Reading materials that exhibit specific properties with respect to a reference model, for example texts about soccer that also contain 10% "new" vocabulary relative to some reference model, can be located quickly. Different methods of selecting texts for reading practice, different constraints on passage characteristics, and different methods of tailoring practice material to the abilities and weaknesses of individual students can be tested.

A particular advantage of this approach to studying lexical mastery is that it enables us to tailor reading practice to the needs of individual students. Statistical language models are ideal for representing the degree to which different words in a large (and growing) vocabulary have been mastered. The main issue in developing and using student-specific language models is acquiring evidence of what a particular student knows at a particular point in time.

1.1 Goals of the system

During the last decade Web search engines have been used to locate texts that satisfy an information need. These engines are designed to run short queries against a huge database of hyperlinked documents very quickly and inexpensively and are carefully tuned for the types of queries that people submit most often (Broder, 2002). A search engine for lexical mastery requires a different approach. An information need in this context might include a topic (e.g., soccer), a reading level (e.g., 4^{th} grade), a list of known vocabulary and the degree to which each word has been mastered, and a constraint on the number of unknown words (e.g., 5%, not counting proper names). Most of this type of information is described naturally by a set of statistical language models. A simple *unigram* language model (a "word histogram") describes the frequency of words/tokens in the sample. *Bigrams* (2 word sequences) and *n-grams* (sequences of *n* words) also exist, but for ad-hoc information retrieval simple unigram language models have been as effective as more complex models (Miller et al., 1999; Ponte and Croft, 1998). When information needs and documents are each represented by statistical language models, documents can be ranked by the similarity of the two language models, i.e., the extent to which they agree on vocabulary and frequency (Ogilvie and Callan, 2002a; Zhai and Lafferty, 2001).

1.1.1 Characterizing the Reading Difficulty of Documents

Reading comprehension is facilitated when documents are at the right reading level, hence the reading difficulty of each document must be determined. There are two facets to reading difficulty for our task: i) the general difficulty of the document, e.g., it contains concepts and vocabulary a 3rd grade student would be expected to understand, and ii) the difficulty for a particular student, e.g., it contains words that this student has mastered. The former, student-independent, analysis can be done as soon as the document is added to the database.

Reading difficulty is a well-studied topic with a rich literature. Many of the well-known methods are crude, for example, based on a linear function of the average number of words per sentence, average number of sentences per paragraph, and similar features. These features are easy to compute, but they are known to be inaccurate, especially on Web pages. The most accurate reading difficulty measures are based on the vocabulary contained in a text. For example, the Revised Dale-Chall measure uses a 3,000-word list that 80% of tested fourth-grade students were able to read (Chall and Dale, 1995). The semantic component of the Revised Dale-Chall measure is a function of the percentage of document terms that do not appear on the 3,000 word list. The more sophisticated Lexile score developed by MetaMetrics (Stenner et al., 1988) is based on vocabulary and word frequency information in a 5-million-word corpus of general school content (Carroll et al., 1971).

Collins-Thompson and Callan have recently developed a new approach to estimating reading difficulty that uses multiple statistical language models (Collins-Thompson and Callan, 2004). Statistical language models can be trained automatically from labelled training data, making them easy to apply to many situations. Different models are created for each level of reading difficulty, e.g., for each grade from kindergarten to 12th grade. Their experiments show that models trained with small amounts of self-labelled Web pages (e.g., pages selected by teachers for 5th grade students, or pages written by 5th grade students) are at least as accurate at predicting the reading difficulty of Web pages as the Revised Dale-Chall, Lexile, or Flesch-Kincaid reading difficulty measures.

1.2 Reader-Specific Lexical Practice for Improved Reading Comprehension

In past (unpublished) work by Eskenazi and Pelton on the representation of the student in NativeAccent^{TM}, a product to teach English pronunciation, the student is characterized at the outset with a set of predefined skills, the level of each determined by a pretest. This information drives a tutoring module to find appropriate material, at the level of the student. As the student uses the tutor, the model of the student (i.e., of what the student knows) is updated. This enables the system to advance to new material chosen specifically for the individual student. The system using this module has been described in (Eskenazi and Hansma, 1998).

1.3 Using the System in an L2 Learning Context

The framework for an L2 reading system is basically the same as the one for an L1 system - student knowledge is represented as a histogram of words that informs the search for new texts. The curriculum can be modelled as another histogram of words, but there is a difference in grain that needs to be accounted for. In L1 learning, we can model US grades 1 to 12, but in L2 learning, there are generally five levels that are defined by the Defense Language Institute and others (for example, the ILR Scale). This would give us a coarser grain of level definition and much more vocabulary to master per level. It would be more desirable to divide the present five levels into "semesters", or half-levels, for finer vocabulary control.

2 The REAP system

In this section, we detail the architecture and retrieval method of the REAP system.

2.1 Architecture

In order for REAP to furnish useful texts, we need a source of authentic documents and a method for retrieving texts appropriate for the individual reader's skill level. The Web is the best source of authentic documents. It provides a large, diverse corpus, composed of exactly the types of texts L2 learners would want to be able to read. In addition, because the corpus would be quite large, we can use stricter criteria in choosing the documents to present to the reader. This allows us more flexibility in terms of how specific our lexical constraints can be, as well as the flexibility to add additional criteria later.

The primary criteria for determining what documents to retrieve are the reader's knowledge of vocabulary and the desired vocabulary knowledge at various points throughout the curriculum. Previous research in K-12 reading difficulty estimation has shown that accuracy and robustness are most improved over traditional measures by using a different model for each grade (Collins-Thompson and Callan, 2004). Thus the curriculum is broken up into levels, where each level is represented as a histogram of words. When building models for either L1 or L2 curriculum, we can learn the level models from a corpus of text that the student would normally read in their studies. This is done automatically, making this system easily trainable for different student populations with different goals.

A student's knowledge can also be modelled using word histograms. In REAP, each student is actually represented by two word histograms, a passive and an active model. The passive model consists of all the words the student has read using our system, along with word frequencies. This can be viewed as an exposure model, where the model is composed of all texts the student has been exposed to while using our system. The active model, on the other hand, includes only the words for which the student has demonstrated knowledge. Ways of demonstrating knowledge are described in more detail later, although the primary way is by answering a question about the word correctly after reading the text. Both the active and passive models are updated each time the student reads a document and is quizzed on the document's new words.

In REAP, we actually use an extended version of the word histograms described above. Each word is also marked with its part of speech; any word with multiple parts of speech is considered to be multiple words. Thus "bat" is treated as two words, a noun and a verb. It is therefore possible to know which meaning of a word a student does or doesn't know. We also removed certain named entities from the curriculum models: person names, organization names, products, works of art, etc. reasoning that these words are not important for comprehension in general, but are elements of the original text that are necessary for comprehension of that specific text. Thus documents with these words will not be specifically sought out to fulfil learning criteria.

We used the Brill tagger to tag parts of speech and IdentiFinder to tag named entities (Brill, 1992; Bikel et al., 1999). Parts of speech and named entities are also tagged in the web corpus. Figure 1 shows the entire REAP system.

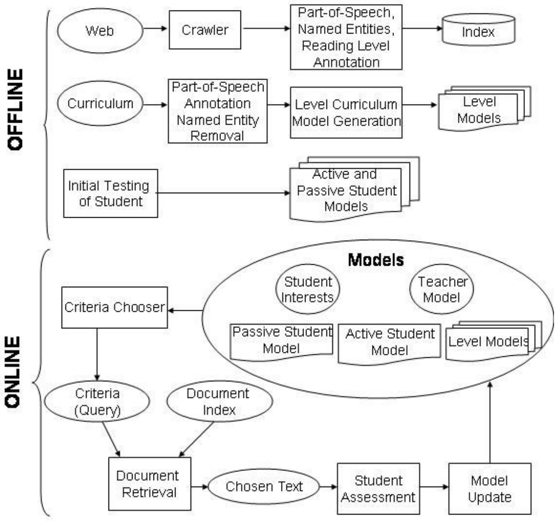

Figure 1. The REAP System.

2.2 Retrieval

To retrieve a new document, REAP's first step is to retrieve the set of documents that match a student's level. To do this, we must know what words the student knows and what words he/she doesn't know. For a given word, the student's passive model can tell us the number of times the student has seen the word, and the active model can tell us the number of times the student has demonstrated knowledge of the word. One method for determining whether the student knows this word is by looking at the number of times the student has demonstrated knowledge of it. For instance, if the student demonstrates knowledge of the word three times in a row in different contexts, we might assume the word is known. Other methods such as ERP, eye-tracking, and analysis of student writings could also be used to determine whether a student knows a word. Regardless of the source, a listing of words that the student knows must be created. At this time we are simply using performance on quizzes designed to test knowledge of the new words in the text.

Given the words a student knows, it is then possible to find documents in the corpus that include some subset of these words, as well as some percentage of new words. This percentage of new words will be the desired vocabulary stretch. When used in a language classroom, this percentage could be the same for everyone or different for individual students. When used in learning experiments, it can be manipulated to determine the effects of different stretch sizes. Retrieval of these documents is done using an extended version of the Lemur Toolkit for Language Modeling and Information Retrieval (Ogilvie and Callan, 2002b).

Now that the system has a set of documents appropriate for the student's reading level, these documents can be ranked according to other criteria. For instance, we can assign a ranking to the words the student does not yet know but that occur in the next proficiency level the student should achieve, based on the curriculum level models. This ranking could be based upon the words' frequencies in the curriculum, so that texts with the most frequent unknown words will be weighted higher than others. Alternatively, we can assign a ranking to the words so that both the most frequent and least frequent words are prioritised. This could be done so that at the end of the semester or year, we are not left with only infrequently occurring words, which will not co-occur often, increasing the number of documents necessary to cover the tail end of the curriculum. The latter method is the default in REAP. In addition to ranking documents in this way, we may also wish to focus on words that the student has seen a number of times but has not yet demonstrated knowledge of. These words can be found by looking at the differences between the student's active and passive models. REAP attempts to focus on documents with these words first, in an attempt to correct deficiencies in the student's knowledge via remediation, and then goes on to documents with never-before-seen words

The system can also make use of teacher input. For instance, the teacher may want the class to read texts about a topic, like the life of George Washington. By building an additional word histogram model for this topic (based on a few documents known to fit the topic), documents can be re-ranked by this model instead of the curriculum model. In this way, we can find documents that are at each individual student's reading level, and are about the topic decided by the teacher. Similarly, word histograms could be built for an upcoming test from specimens of past tests, so that students could gain additional exposure to words that they will soon be tested on. If these models are weighted, it is also possible for a teacher to specify the importance of each. A teacher is then able to spend certain class periods focused on particular topics, certain class periods focused on improving vocabulary, etc. REAP can also make use of student interests in the same way as teacher topics. In parallel to improving vocabulary, the system will prefer documents that satisfy student interests. As mentioned earlier, because of the size of the web, the system has greater flexibility when trying to find documents that match multiple criteria. The larger the corpus, the greater the flexibility.

3 Conclusion

We have described the REAP system, which furnishes appropriate authentic texts for L1 or L2 reading. The system is being expanded to include such other elements as grammar difficulty and document cohesiveness.

Acknowledgements

The authors would like to thank Jamie Callan and Kevyn Collins-Thompson for their help in this research.

This project is supported by Award # R305G030123 Funded from the U.S. Department of Education, Institute of Education Sciences. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Department of Education.

References

A. Biemiller and N. Stonim. 2001. Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition, *Journal of Educational Psychology*, 93 (3): 498-520.

D. Bikel, R. Schwartz, and R. Weischedel. 1999. An Algorithm that Learns What's in a Name, *Machine Learning Journal Special Issue on Natural Language Learning*, 34: 211-231.

E. Brill. 1992. A Simple Rule-Based Part of Speech Tagger, *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, 152-155.

A. Broder. 2002. A taxonomy of Web search, *SIGIR Forum*, 36 (2), 3-10.

J.B. Carroll, P. Davies, and B. Richman. 1971. *Word frequency book*. Boston: Houghton Mifflin.

J.S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula,* Brookline Books, Cambridge, MA.

K. Collins-Thompson and J. Callan. 2004. A language modeling approach to predicting reading difficulty, *Proceedings of the HLT/NAACL 2004 Conference*, Boston.

M. Eskenazi and S. Hansma. 1998. The Fluency Pronunciation Trainer, *Proc. STiLL Workshop on Speech Technology in Language Learning*, Marhollmen, May, 1998.

ILR scale - The Interagency Language Roundtable Scale, The Globe-Gate Project http://www.utm.edu/~globeg/ilrhome.shtml

D.R.H. Miller, T. Leek, and R.M. Schwartz. 1999. A Hidden Markov Model information retrieval system. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

P. Ogilvie and J. Callan. 2002. Language models and structured documents. *INEX 2002 Workshop Proceedings,* 18-23.

P. Ogilvie and J. Callan. 2002. Experiments using the Lemur toolkit, *Proceedings of the 2001 Text REtrieval Conference*, NIST special publication 500-250: 103-108.

C.A. Perfetti, M.A. Britt, and M. Georgi. 1995. *Text-based learning and reasoning: Studies in history*. Hillsdale, NJ: Erlbaum, p. 28.

C.A. Perfetti and L. Hart. 2001. The lexical quality hypothesis, In L. Vehoeven. C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy*, Amsterdam/Philadelphia: John Benjamins.

J. Ponte and B. Croft. 1998. A Language Modeling Approach to Information Retrieval. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 275-281.

A.J. Stenner, I. Horabin, D.R. Smith, and M. Smith. 1988. *The Lexile framework*, Durham, NC: Metametrics.

C. Zhai and J. Lafferty. 2001. A study of smoothing methods in language models applied to adhoc information retrieval. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.