STUDENT, TEXT AND CURRICULUM MODELING FOR READER-SPECIFIC DOCUMENT RETRIEVAL

Jonathan C. Brown and Maxine Eskenazi
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA
USA

jonbrown@cs.cmu.edu
max@cs.cmu.edu

ABSTRACT
In today's language-learning classrooms, all of the students in a class almost always have the same text to read. Although students have different reading levels, it is impractical for a single teacher to find unique texts matched to each student's abilities. The REAP system was developed to make the process of providing students with individualized texts practical. The texts come in the form of authentic documents retrieved from the Web, and the system tracks and assesses students' knowledge as they use the system. The system is able to find documents that meet various and individualized criteria. In this paper, we describe our work on modeling lexical familiarity. In particular, we detail the approaches taken for modeling the student's vocabulary knowledge, the contents of documents in the corpus, and the components of the curriculum. We also address related work and future plans.

KEY WORDS
User modeling and adaptation, computer-based learning, language-learning, and reading-level personalization.

1. Introduction

The REAP project in the CMU Language Technologies Institute has created software that is designed to find appropriate authentic documents for a student who is either learning to read either their native or a new language. In this paper we will focus on the part of the project that focuses on modeling lexical familiarity. This includes modeling a student's reading level, the reading texts themselves, and finally the curriculum.

In today's classrooms, students often read prepared texts that are not authentic for practice. There are two disadvantages to this. First, there is increasing realization that it is import to use authentic documents to teach reading. That is, when reading prepared texts, students are not being exposed to examples of the real language that is used in everyday written communication. Second, the students all get the same text to read. Students who are having trouble learning certain vocabulary items have little chance for remediation. Similarly, students who are ahead of others do not have the opportunity to advance more quickly. Mastery of core vocabulary at a particular reading level is essential to the development of more complex reading comprehension skills [1][2]. In addition, it has been shown that 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 [3]. The REAP system supplies texts that are both authentic and personalized to the individual user's reading-level. This allows students to acquire the necessary lexical mastery as quickly as possible, without passing some students by or requiring other students to slow down.

The goal of the REAP project is twofold. First, we aim to create a framework that presents individual students with texts matched to their own reading levels. This can be used in the classroom, where the system becomes an extension of teacher time. Second, we aim to enhance the ability of learning researchers to test hypotheses on how to improve vocabulary skills for L1 (first-language) and L2 (second-language) learners. That is, the system is designed to use very specific criteria for the types of documents that are retrieved, so that researchers can conduct controlled experiments to study, for example, how well students learn vocabulary from context, the optimal percentage of unknown words to present within a document, or the most useful types of help to provide students.

2. The Architecture of REAP

Figure 1 shows the components of the REAP system. The top half of the figure shows the processes which work offline, before the system is ready to be used. The bottom half shows the online processes.

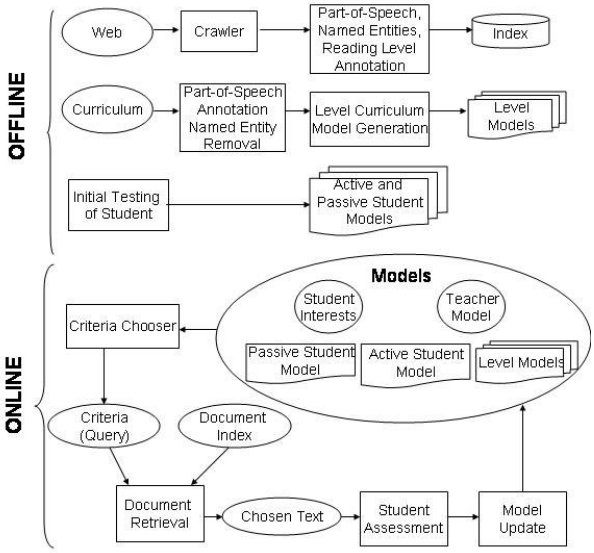

Figure 1. The REAP System.

The online processes consist of three separate tasks. First, an index must be built. This index will consist of the entire document corpus from which we will retrieve texts. The current size of this corpus is approximately 10 million documents, with a goal of at least 20 million pages consisting of material for grades 1 through 8 [4]. This material is procured from the Web, which 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 second task is that the curriculum model must be created. Finally, the student model must be instantiated to the starting knowledge state of the user. The latter two processes are described in sections 3 and 6.

The online portion of the system consists of three main components. The first component is the model collection. This component houses of all of the data necessary for modeling the student and the curriculum. This will be described in detail in sections 4 and 5. The second component is the retrieval engine. Using data from the models, as well as any additional criteria specified by a teacher or researcher, the system is able to find documents that match these criteria and are personalized to the student. Finally, there is an assessment and update component. After a student has read a document, that student will be assessed on the new vocabulary terms that appeared in the document. Then, the student model will be updated to indicate the new state of knowledge. At this time, the system is able to find the next document for the user, based on either the same or new criteria.

3. Modeling the Curriculum

In order to model an entire curriculum that covers multiple grades, we break it up into levels. Each level is represented as a histogram of words. A word histogram is basically a simple statistical language model. Statistical language models have been used successfully in speech recognition, information retrieval, and other language technologies [5][6]. Each level of the curriculum is then composed of a list of words and their frequencies within the texts of that level.

We can build these models for either an L1 or L2 curriculum by using a corpus of texts that the students would normally read in their studies. This way, we can represent the frequencies of the words at each level in the curriculum and determine at which level of the curriculum a given word is introduced. This is all done automatically. This enables the system to be easily trained for different student populations with different goals.

One issue that arises when dealing with individual words is that of word ambiguity. For instance, if the word “bat” appeared in one of the texts used in training the curriculum model, the word may refer to the flying animal, to the wooden piece of sports equipment, or to the action of batting one’s eyelashes. Without some way of performing disambiguation, we will not be able to specify which sense of the word we intend to be part of the curriculum and thus learned by the students. REAP reduces the impact of this issue by modeling part of speech information. That is, we use an extended version of word histograms where each word is annotated with its parts of speech (POS). For any word with multiple parts of speech, only those word/POS combinations which appear in the training materials are included in the curriculum models. For instance, if the word “bat” appears in the training materials as both a noun and a verb, there will be two entries in the level model. Note that this allows us to know whether the students are expected to learn the noun form, verb form, or both.

However, this does not allow us to determine which of the two noun forms of the word "bat" is intended. Thus, the system finds documents that use the word in any of its noun senses.

Another issue concerns named entities. Named entities are person names, organization names, product names, works of art, and so on. Although these words routinely appear in the texts used to learn the curriculum models, they are almost never important words which we want the system to seek out for the user to learn. Thus, we remove these words from the curriculum models so that the system will not actively seek out documents in which these terms appear. We use the Brill tagger to tag part of speech information and IndentiFinder to tag named entities [7][8].

4. Modeling the Students and Documents

In order to match students and documents, we must characterize the reading level of the documents in the corpus and the reading ability of the user. These two measures must be reconcilable in some way in order to perform the retrieval process. Because of this, the modeling of the corpus and the modeling of the user are heavily dependent upon one another and will be discussed together.

The simplest method of characterizing a student or a document is by using a single number. This number often corresponds to a value called readability. Readability is used to describe how easy or difficult a given text is to read. There are numerous characteristics that affect the readability of documents. For instance, a document can be levelled in terms of the difficulty of the discourse structure, the difficulty of specific grammatical structures used, or the number of abstract or otherwise difficult terms in the document. A number of formulas exist to measure readability. Because of the difficulty in measuring items like discourse structure, most of these measures rely on much simpler features, such as the average sentence length or the average word length. These are clearly surface features, although they have been found to correlate well with the deeper characteristics.

Three of the most common measures used are the FOG [9], SMOG [10], and Flesch-Kincaid [11] metrics. All three of these measures take a text as input and produce a single number as output. This number corresponds to the estimated grade level for a document, or the years of schooling one would be expected to need to have completed in order to read the text. As an example, consider the Flesch-Kincaid measure, a U.S. Department of Defense standard that was developed to test the readability of military training manuals. It is defined as .39 * the average number of words per sentence + 11.80 * the average number of syllables per word – 15.59 [11]. Although these methods are easy to compute, different readability measures often give widely varying results for the same text. Both the designers of these methods [12] and other researchers [13] have pointed this out, and recommend that these measures only be used as rough guides.

More 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 [14]. While the syntactic component of this measure is similar to the three simple measures described above, this measure also has a semantic component based on the percentage of terms in the document which do not appear in the 3,000-word list. The Lexile measure [15], developed by MetaMetrics, is a similar, but more sophisticated, method. It is based on the mean log frequency of the text based on vocabulary and word frequency data from a 5-million-word corpus of general school content [16].

One disadvantage of measures such as Revised Dale-Chall and Lexile is that they assume the text being evaluated is at least 100 words long and is composed of well-defined sentences. Neither of these assumptions is always true for the Web documents that compose the REAP corpus. Consequently, other measures have been defined for Web and other non-traditional documents [17][18]. One of these is a readability measure that uses multiple statistical language models that has been developed recently by Collins-Thompson and Callan [18]. Statistical language models can be trained automatically

from labelled training data. Different models are created for each level of reading difficulty, e.g. for each grade from kindergarten to 12th grade. These models can also be smoothed to account for noise in the data and to reduce data requirements. Collins-Thompson and Callan trained their models 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), and compared the resulting grade-level classifier to the Flesch-Kincaid measure as well as a number of commonly-used semantic measures such as mean log frequency and percentage of unknown tokens given a familiar-word list. The language modeling based classifier outperforms the other measures for web documents and very short documents. For this reason, we use this classifier in REAP to assign grade-levels to all the documents in the corpus.

Given a readability measure such as the language modeling classifier of Collins-Thompson and Callan, one can assign a grade-level to every document, and retrieve documents that are annotated with the current grade level of the student. For many tasks this may be sufficient. However, for the types of lexical constraints the REAP system must satisfy, finding documents at the correct grade-level is only part of the problem. For instance, one of the primary criteria is to find documents with a given percentage of new words. That is, to find documents were some percentage of the words are believed to be known by the student, and the remaining words are words which are in the next level of the curriculum model described in section 3 of this paper. Clearly, to satisfy criteria such as this, the system must have a model of the current state of knowledge of the student that is at the word level.

A student's word knowledge can also be modeled 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 is basically an exposure model, where the model consists of all texts the student has read using our system. The active model, however, consists only of the words for which the student has demonstrated knowledge. See the section on procuring data for the user model to see how REAP handles this. Both the active and passive models are updated each time the student reads a document and demonstrates knowledge of specific new words which occurred in the document.

Just as in curriculum modeling, the issue of word ambiguity arises in student modeling. We need to be able to represent which sense of the word the student knows or doesn't know. We use the same technique of annotating words with part of speech information, in order to deal with this issue. Of course, this does not allow us to distinguish between two noun senses of a word. We plan to employ more advanced word sense disambiguation techniques in the near future, in order to map uses of words in context to their senses in WordNet [19].

5. Modeling Other Aspects

In addition to the curriculum, the system must also be able to make use of both teacher input and student interests in retrieving documents. Because of this, we must be able to model both of these as well. For instance, a teacher may want the class to read texts about a specific topic, like the life of George Washington. By building an additional word histogram model for this topic, documents can be re-ranked by this model instead of the curriculum model. REAP is thus able to find documents at each individual student's reading level and about the topic at hand. Topic models such as these can be built on the fly, based on a few documents that are known to fit the topic. Similarly, word histograms could be built for specific student interests, based on documents the students have chosen as being interesting to them. The REAP system is able to use multiple criteria at one time as well. For more details on this and on the retrieval process in general, see [20]. The larger the corpus of documents, the more detailed retrieval criteria can become.

6. Procuring Data for the Student Model

In a system where one uses a readability measure of some kind to annotate each document with a single number, the student model only needs to use one number as well.

When this corresponds to grade level, this number is especially easy to get. However, when modeling the specific vocabulary knowledge of the student, as REAP does, you clearly need more. Specifically, you must have a method for initializing the student model as well as for continually updating that student model.

Just as it is not feasible for a teacher to find individual texts for each student in a class, it is also not feasible to require the teacher to create questions for each vocabulary word. Therefore, we take the approach of automatically generating questions to assess students both during student model initialization and after every text with new vocabulary terms that a student reads. These computer-generated questions test various aspects of word knowledge, such definitional knowledge or whether a student can use a word in context. A study is underway at the time of this writing to compare these questions to human-generated ones.

7. Related Work

One system that has a goal similar to that of REAP is the Squirrel system [21]. The goal of this system is to retrieve texts for students to read targeted at a specific reading level. The target population is second language learners of Nordic languages. The system is invoked by an example text provided by the user. The Squirrel system then analyses this text and assigns it a score, based on the Lix readability formula [22]. The Lix formula is similar in nature to the simple readability formulas like FOG and SMOG mentioned earlier, although it has been used more than other measures for computing readability scores for the target languages of this system. After computing the Lix score for the example document, the system then finds documents which are of a similar topic and Lix score to the example. The authors of this study do not report how well the Lix measure works for Web documents. This would be worth exploring further. One advantage for such a system over REAP is that, because the Lix measure can easily be computed on the fly, the system can be, and actually was, implemented as a meta-search system, thus eliminating the need for their own Web corpus. The system can be built on top of an existing search engine. Of course, this limits the modeling that can be done to measures which can be computed on the fly, which would rule out REAP's finer vocabulary method as well as our future system which will use more complex features.

Another related system is the IWiLL system. One component of this system is the Lexical Difficulty Filter [23]. The filter uses a word frequency list and a user-defined frequency threshold to filter out text where some percentage of the words fall outside the frequency threshold. The system finds documents in a similar way to REAP, where some percentage of the terms in the document must be known to the user. However, instead of modeling full-vocabulary knowledge, the system works under the assumption that the state of the user's knowledge can be represented as a boundary, where all words to one side occur more frequently than words to the other side. One assumes that the user knows all of the frequent words and none of the unfrequent words. It remains to be seen whether someone's vocabulary knowledge can be accurately represented in this way.

Liu, et al, describe a system for implicitly inferring a user's reading level using a user's query history [24]. This is an interesting approach to determine the user's reading ability. The authors point out the poor performance of readability measures such as FOG, SMOG, and Flesch-Kincaid on extremely short passages, such as the fewer than ten word queries they focus on. Consequently, they implement their own measure, using both semantic and syntactic features and an SVM (Support Vector Machines) classifier. They were able to classify queries into grade-level categories relatively well, while FOG, SMOG, and Flesch-Kincaid all performed worse than random selection. This approach could be useful for cases when the full-vocabulary knowledge method of REAP is unnecessary.

8. Future Work

There are a number of elements of this project in which we have ongoing extensions. One of these areas is extending the student model to incorporate, along with the active and passive models, a prediction model. This model will represent the estimated state of knowledge of

the user based not only on what knowledge they have explicitly demonstrated but also on our predictions of knowledge on untested words. These predictions will be based on at least two things, word cohorts and student confidence ratings.

Another area is in moving beyond vocabulary to model more complex features. Both the language modeling grade-level classifier of Collins-Thompson and Callan and the finer full-vocabulary knowledge method described in this paper rely almost completely on vocabulary. Thus, they are missing the syntactic components of the traditional readability measures. Although we have seen that simple syntactic features, such as the average number of syllables per word, are ineffective features for grade-level annotation of Web documents, we should keep in mind that these features were used in an effort to correlate with deeper features that were more difficult to measure, such as discourse structure of the text or the difficulty of the grammatical structures used in the text. We reason that these deeper features could still be useful for characterizing Web documents, and that is was simply the case that the simpler features used in place of them in previous readability measures were not sufficient for these types of documents. There may be other simple features that correlate better with these deeper features for Web documents, or we may need to attempt to model the deeper features directly. We also plan to allow constraints on other difficulty measures such s text cohesion, text coherence, and discourse structure [25].

Finally, it would be interesting to apply techniques such as those used in REAP in a standard information retrieval system. In [24], researchers show that document relevance is determined not only by topic-relevance but also by level-relevance. Documents should not be considered relevant unless they are of the same topic as the query and of the correct user reading level. Reading-level personalization should improve system performance.

9. Conclusion

This paper has described REAP and detailed the techniques used in student, document, and curriculum modeling. REAP can be used both as an in-classroom resource and as a platform for learning experiments. The system is able to find documents with specific lexical constraints, and in the future will also employ grammatically analysis and other measures for use as constraints on retrieval. We believe REAP will be very useful for learning experiments, some of which have recently begun. We also believe that REAP furthers the state of the art in reading-level personalization.

10. 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:

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

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

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

[4] K. Collins-Thompson and J. Callan. Information retrieval for language tutoring: An overview of the REAP project (poster description). In *Proceedings of the Twenty Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Sheffield, UK, 2004.

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

[6] R. Rosenfeld. Two decades of Statistical Language Modeling: Where Do We Go From Here? P*roceedings of the IEEE*, 88(8), 2000.

[7] E. Brill. A Simple Rule-Based Part of Speech Tagger, *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, 1992, 152-155.
[8] D. Bikel, R. Schwartz, and R. Weischedel. An Algorithm that Learns What's in a Name, *Machine Learning Journal Special Issue on Natural Language Learning*, 34, 1999: 211-231.
[9] R. Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.
[10] H. McLaughlin. SMOG grading – a new readability formula. *Journal of Reading*, 22, 1962.
[11] J. Kincaid, R. Fishburne, R. Rodgers and B. Chissom. Derivation of new readability formulas for navy enlisted personnel. *Branch Report 8-75*. Millington, TN: Chief of Naval Training, 1975.
[12] G. R. Klare, P. P. Rowe, M. G. St. John, and L. M. Stolurow. Automation of the Flesch reading ease readability formula, with various options. *Reading Research Quarterly 4*, 1969.
[13] W. H. DuBay. The Principles of Readability. Costa Mesa, CA: Impact Information., 2004.
[14] J.S. Chall and E. Dale. *Readability Revisited: The New Dale-Chall Readability Formula,* Brookline Books, Cambridge, MA, 1995.
[15] A.J. Stenner, I. Horabin, D.R. Smith, and M. Smith. *The Lexile framework*, Durham, NC: Metametrics, 1988.
[16] J.B. Carroll, P. Davies, and B. Richman. *Word frequency book*. Boston: Houghton Mifflin, 1971.
[17] L. Si and J. Callan. A statistical model for scientific readability (poster description). In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*. ACM, 2001.
[18] K. Collins-Thompson and J. Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of the HLT/NAACL 2004 Conference*. Boston, 2004.
[19] WordNet. http://wordnet.princeton.edu/
[20] J. Brown and M. Eskenazi. Retrieval of Authentic Documents for Reader-Specific Lexical Practice. In *Proceedings of InSTIL/ICALL Symposium 2004*. Venice, Italy, 2004.
[21] K. Nilsson, and L. Borin. Living off the land: The Web as a source of practice texts for learners of less prevalent languages. *Proceedings of LREC 2002*, Third International Conference on Language Resources and Evaluation, Las Palmas: ELRA, 2002.
[22] C. H. Björnsson. *Läsbarhet*. Lund: Liber, 1968.
[23] D. Wible, F. Chien, C. Kuo, and T.C. Wang. Adjusting corpus searches for learners' level: filtering results for frequency. Presentation at *TALC 2000*, Graz, Austria, 2000.
[24] X. Liu, W. B. Croft, P. Oh, and D. Hart. Automatic Recognition of Reading Levels from User Queries. *SIGIR '04*. Sheffield, England, 2004.
[25] D. S. McNamara, M. M. Louwerse and A. C. Graesser. Unpublished. Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Grant proposal.