Automating Repeated Exposure to Target Vocabulary for Second Language Learners

David Wible*, Chin-Hwa Kuo**, Feng-yi Chien*, Nai Lung Taso**,
Graduate Institute of Western Languages and Literature, Tamkang University,
dwible@mail.tku.edu.tw
Computers and Networking (CAN) Laboratory, Department of Computer Science and Information Engineering, Tamkang University, **chkuo@mail.tku.edu.tw**

Abstract

A web-based tool is described, called the Supplementary Reading Provider (SRP), devised to find related readings according to learners' lexical level which offer repeated exposure to target vocabulary. The SRP exploits text retrieval techniques based upon the hypothesis that there is a parallel between text similarity measurement on the one hand and the pedagogical task of providing supplementary readings which offer repeated exposure to new vocabulary on the other. Two criteria are compared for finding supplementary passages from a corpus of reading: one matches keywords of the original passage and those of the corpus passages, the other matches target vocabulary words without regard to keywords. The integration of SRP into an web-based language learning platform called IWiLL is described as well.

1. Background

1.1 Vocabulary Acquisition and Input

Research on reading and on vocabulary acquisition suggest the value of providing supplementary readings which offer learners repeated exposure to target vocabulary in context. Nation (1990) points out "...the effort given to the learning of new words will be wasted if this is not followed up by later meeting with the words." Thus, the "...increase in vocabulary size must be accompanied by many opportunities to put this vocabulary to use." (p.119) It is just such opportunities which are lacking in much high school EFL curricula under the pressures of time and the influence of vocabulary teaching traditions based upon rote memorization.

It is clear that the similarity of the supplementary readings and the repetition of target words in context form a meaningful reading for learners on the assumption that "...only contexts will fully demonstrate the semantic, syntactic, and collocational features of a word the learner has to process in order to establish the numerous links and associations with other words necessary for easy accessibility and retrieval" (Groot 2000. p.65).

2. Setting

The tool, called Supplementary Reading Provider (SRP), is designed to be one component in a larger integrated web-based language learning environment, IWiLL. IWiLL consists of several highly integrated components that support language learners, teachers, and researchers. The most mature component is designed to support English composition. A more recent module is the reading component. VoD (video on demand) tools designed by the IWiLL team are also integrated into the environment, allowing teachers and learners to do selective searches for specific sorts of linguistic input for the learners. SRP is intended as one tool within this platform. Currently the SRP is a stand-alone tool that provides teachers and student search capabilities for supplementary readings online or local reading search results. This is an initial stage in our development of the SRP. The SRP will have a web-based version whereby users need only a common commercial browser such as Internet Explorer or Netscape, and the tool will be accessible from the web. More specifically, we will incorporate it into a suite of tools available on our web-based language learning environment (IWiLL).

3. Method

To describe the method used in SRP, it will be helpful to imagine a particular setting. We assume a learner who is reading a lesson online in a web-based learning platform (e.g., IWiLL) and the passage has an accompanying list of target vocabulary items that appear in the passage. We will refer to the vocabulary token in the textbook passage as the target vocabulary item. The system will search a corpus of texts which we will simply refer to as the corpus. The aim is for the SRP to take a target vocabulary item as input and provide as output a set of texts from the corpus that contain tokens of the target vocabulary which resemble the original semantically and of course match it in part of speech.

We compare two algorithms for this search: one which relies on matching the keywords of the original textbook text and those of the texts in the corpus and

0-7695-1013-2/01 $10.00 © 2001 IEEE

another which matches the vocabulary list from the textbook text to the texts in the corpus rather than keywords.

For the purposes of testing the precision of this search technique, we use a sense tagged Brown corpus bundled in WordNet 1.6. One of the texts in the Brown corpus is selected as the "textbook text" and the remainder of the corpus serves as the corpus. The search technique makes no reference to the semantic tags, but we use those tags post hoc to evaluate the accuracy of the search results. If the semantic tag (the synset index) of the textbook vocabulary token matches the semantic tag of a matching token retrieved by the algorithm, we consider this a hit, that is, an accurate retrieval that would give the learner repeated exposure to the same target vocabulary word with the same sense encountered in the textbook.

In order to provide "similar" readings with repeated vocabulary items from high school coursebook materials, we use text retrieval technique to calculate the similarity between coursebook materials and supplemental readings. We use VSM (Vector space model) to build the feature vector of each document. Every document in teaching materials has two kinds of feature vectors: one is composed of a vocabulary list and the other is composed of keywords extracted automatically by the system. However, the supplemental readings only encode the second feature vector. Thus, we use a frequency list file to represent the feature vector. It is exemplified in the following list:

learning, 2
english, 8
easy, 1
student, 1
asks, 1
questions, 1

The first element is the keyword or vocabulary items and the second is the frequency of the word in the document. The feature vector composed by keywords is automatically extracted by the system. After extracting every single word in the document, the system will delete stop words from feature vector.

After the document preprocessing, we use a cosine measure to calculate the similarity between the textbook passage and each of the supplemental readings, first according to the vector calculated by keywords of the textbook passage and second according to the vector calculated by target vocabulary items in the textbook passage . The cosine formulation is listed below.

$$sim(D_1, D_2) = \frac{\sum d_{1k} \times d_{2k}}{\sqrt{\sum d_{1k}^{\ 2} \times \sum d_{2k}^{\ 2}}}$$

where D_1, D_2 are the two documents, d_{ik} is the k^{th} feature vector value of the D_i. We run these calculations on the Brown Corpus that has been semantically tagged according to WordNet senses. Results of the similarity calculations are compared to the Wordnet sense tags to test the comparative accuracy of using target vocabulary vs keywords as the evaluation standard for determining similarity.

4. Conclusion

Text similarity measurement and text retrieval techniques offer a range of ways to automate the task of providing learners with passages that give repeated exposure to target vocabulary sufficiently similar to the original usage to be helpful to the learner. Two specific approaches are tested. These techniques (and potentially a range of others) can improve upon the low precision of mere string matching KWIC searches for vocabulary yet circumvent the need for relying on heavily annotated, semantically tagged corpora to achieve this precision. Future research is needed to tease out a number of variables for improving this precision and applying the results to second language vocabulary acquisition.

5. References

Barnbrook, Geoff. (1996) *Language and Computers: A Practical Introduction to the Computer Analysis of Language.* Edinburgh: Edinburgh University Press.

Fellbaum, Christiane (1998) *Word Net: An Electronic Lexical Database.* Massachusetts: MIT Press.

Goethals, Michael. (1997) "How useful is word frequency information for the EFL teacher (and/or learner)?" Paper presented at EUROCALL Conference Dublin.

Groot, Peter J.M. (2000) "Computer Assisted Second Language Vocabulary Acquisition." Language Learning & Technology Vol4, No.1, 60-81. On-line. Available from Explore @ http://llt.msu.edu/vo14numl/groot/default.html

Krashen, S. (1995) *The Input Hypothesis: Issues and Implications.* Londons: Longman

Nation, J.S.P. (1990) *Teaching and Learning Vocabulary.* Massachusetts: Heinle &Heinle

Summers, Della. (1995) "Computer lexicography: the importance of representative ness in relation to frequency." In Thomas, Jenny & Short Mick, *Using Corpora for Language Research.*