Application of Automatic Thesaurus Extraction for Computer Generation of Vocabulary Questions

Michael Heilman and Maxine Eskenazi

Language Technologies Institute
Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A.
{mheilman,max}@cs.cmu.edu

Abstract

Automatic thesaurus extraction techniques are applied to computer-generated related word vocabulary questions. These questions assess and provide practice for an aspect of word knowledge found to be important for language learning. Automatic generation of such questions reduces the need for human authoring of practice materials. In evaluations with real teachers, most of the generated questions were considered to be usable in real classrooms. Also, performance of native and non-native speakers on these automatically generated questions was similar to their performance on manually generated questions for the same words. This application of natural language processing techniques to English as a Second language education is a promising step toward automatically producing vocabulary practice and assessment materials for language learners.

Index Terms: automatic thesaurus extraction, Word Associates Test, computer assisted language learning

1. Introduction

The goal of this work is to automatically generate a specific type of vocabulary assessment item, in order to add practice and assessment questions to REAP [1], a tutoring system for vocabulary that is currently used for English as a Second Language (ESL). The REAP system provides ESL learners with individualized and adaptive practice by presenting a series of readings and practice exercises that match their personal learning needs. A goal of the system is to allow learners to study any unknown word that appears in a reading, rather than just a restricted set of words in a list of target vocabulary. Using computer-generated questions enables the REAP system to provide practice for and assess knowledge of potentially any word without requiring teachers to manually author questions. Previous work has shown that other types of vocabulary assessment items can be automatically generated. For example, Brown, et al. [2] developed a system that uses WordNet [3] as a preexisting lexical knowledge base for generating various multiple-choice questions. Additionally, systems have been developed for automatically generating cloze, or fill-in-the-blank, questions from large corpora of text [4].

The present work aims at generating questions similar to the items found in John Read's Word Associates Test [5]. In the Word Associate's Test, students identify related words, called associates, for a given word. These related words are selected from among a set of distractors. Read employs three types of relations. The first type is paradigmatic, where the related words are synonyms, antonyms, or similar but more or less specific in meaning. The second type is syntagmatic, whereby the words frequently collocate, or occur together, in phrases. The third type is analytic, where one word is a part or aspect of the other, but was excluded from later versions of the Word Associates Test. Nation [6] has identified knowledge about the associations of a word as one of the eight features of word knowledge.

This work focuses on generating items that address paradigmatic relations for a target word. The system creates questions such as the one in Figure 1, where a student identifies words that are either near matches, opposites, generalizations, or specific types of the concept represented by a given word.

Figure 1. *An example automatically generated related word question, with the correct answer italicized*

Which set of words are most related in meaning to "reject"?
A. pray, forget, remember
B. invest, total, owe
C. accept, oppose, approve
D. persuade, convince, anger

One possible approach would be to use a manually-created lexical knowledge base such as WordNet [3] or a thesaurus to generate such questions--as in the work by Brown, et al. [2]. Implementation of Brown's work revealed issues associated with using manually generated sources of lexical knowledge. Coverage in manually generated thesauri may be poor for rare words with, for example, few or even no specified synonyms. As noted by Curran and Moens [7], manually generated thesauri may also exhibit bias and inconsistency. For instance, the information about a given word may be biased by an author's personal intuitions about a word's meaning. Perhaps the greatest disadvantage of using manually generated thesauri or lexical resources is that they must be manually authored. Creating a reliable, broad-coverage thesaurus is not a feasible task for a language teacher or creator of a computer-assisted language learning application.

Therefore, we applied an automatic thesaurus extraction technique to generate a knowledge base for generating related word questions. The fundamental idea behind automatic thesaurus extraction is the "distributional hypothesis," which states that "the meaning of entities…is related to the restriction of combinations of these entities relative to other entities" [8]. For example, consider the words "milk," "juice," and "cup." These words might appear in similar contexts about food and meals. However, the first two are likely to be objects of verbs

such as "drink" or "pour," while the word "cup" would not. By their use of such dependency relationships, automatic thesaurus extraction techniques are different from techniques such as Latent Semantic Analysis [9] that simply use the co-occurrence of words in the same passage as the primary means of measuring semantic similarity.

Hindle [10] applied the distributional hypothesis to classify nouns based on distributions of automatically identified subject, verb, and object relations. Others, including Lin [11], have employed various statistical measures for identifying semantically related words based on the distributions of relations. Curran and Moens [7] provide a review of recent automatic thesaurus extraction techniques, as well as an evaluation of a number of different statistical measures and weighting schemes.

2. Applying Automatic Thesaurus Extraction

Our approach to thesaurus extraction is similar to Lin [11]. We use a broad-coverage dependency parser [12] to create dependency triples for all of the dependency relations extracted from a large corpus of text [13]. These triples contain a target word, a dependency, and a dependent word. For example, from the sentence "I drank milk this morning,", the triple "(drink, obj, milk)" would be extracted for the dependency relation indicating that the noun "milk" is the object of the main verb "drank." These triples are grouped by their target words so that the system has counts for each triple in which a given target word occurs. In this paper, (*w, r, w'*) denotes a particular triple, and p(*w, r, w'*) denotes the maximum likelihood estimate of the probability of that triple occurring in text, which is the count for the triple divided by the total number of relations in the corpus.

Using probabilities of triples directly is problematic because some dependent words are more or less likely *a priori* than others. For example, a small number of occurrences of the triple ("dog", modified-by, "ferocious") are much more informative than ("dog", modified-by, "good"), even though the latter might occur more frequently. We used mutual information to normalize the counts of triples, although other statistics such as X^2 have been employed. Relations with negative mutual information values are ignored. The mutual information for a given triple is defined as in Equation (1), where * indicates the union of relation types or words. For example, p(*w,*,**) indicates the overall estimated probability of word *w* occurring, in any relation type (e.g., object-of) with any other word.

$$MI = \log \frac{p(w,r,w')}{p(w,*,*)\,p(*,r,w')} \quad (1)$$

The process of identifying informative dependency triples is, in fact, very similar to the process of collocation extraction [14], in which lexical preferences are identified. Computing the informativeness of each dependency triple produces a distribution, or feature set, of lexical preferences for a given word. Thus, we expect to find that the relation "ferocious dog" is informative, the relation "good dog" is less informative, and that a relation that is unlikely to occur, such as "paper dog", has almost no value at all. The next step is to compute similarities for each word pair based on these lexical preferences. For example, to find out what is related in meaning to "dog," the system identifies what else can be described as "ferocious," and what other words can be the object of the verb "to pet."

Curran and Moens [7] discuss and evaluate a number of measures to compute similarity from the distribution of lexical preferences. We implemented the two most successful measures: Lin's measure [11], and the Jaccard measure employed by Grefenstette [15]. We finally retained the Jaccard measure because it performed slightly better overall in our evaluations, which agrees with Curran and Moen's results.

To compute the similarity or relatedness of two words, w_1 and w_2, using the Jaccard measure, the system first finds the union of informative relations for those words. If the informativeness weight for a relation is defined as wgt(*w, r, w'*), then this set consists of pairs (*r, w'*) for which either wgt(w_1, *r, w'*) or wgt(w_2, *r, w'*) is positive. One of these weights will be less than or equal to the other. The similarity of two words is defined as the sum, over all pairs, of the lower weight divided by the higher weight, as in Equation (2).

$$\mathrm{Sim}_{\mathrm{JACCARD}}(w_1, w_2) = \sum_{(r,w')} \frac{\min(wgt(w_1,r,w'), wgt(w_2,r,w'))}{\max(wgt(w_1,r,w'), wgt(w_2,r,w'))} \quad (2)$$

The first step in applying the above measures was to parse a corpus of text and extract dependency triples. A broad coverage parser [12] was used to extract triples from a subset of the Gigaword corpus [13] consisting of approximately 500 million words. The one-time process of parsing the text took approximately one week. The system then automatically extracted sets of related words from a corpus of text by using the similarity and weighting functions described above. The system extracted 100 related words for each of 20,000 head words. Some examples are shown in Table 1.

Table 1. *Example related words. The target word is shown at the top of each column in bold.*

adequate	**bias**	**enhance**
sufficient	prejudice	boost
inadequate	discrimination	strengthen
proper	racism	ensure
insufficient	animosity	improve
appropriate	interference	expand
reasonable	shortcoming	bolster
necessary	imbalance	achieve
minimal	hostility	promote
needed	ignorance	restore
satisfactory	perception	guarantee

There are some practical considerations in the application of thesaurus extraction to assess and provide practice for language learners, and so the vocabulary was restricted. For example, a related word question for the word "sad" should not contain the word "morose" because the learner's knowledge of the rarer word, "morose," is not what is being tested. Thus, words that were either a) not in a dictionary, or b) below a certain frequency threshold in the corpus were ignored.

Once the thesaurus is extracted and sets of related words are defined, the creation of related word assessment items is fairly straightforward. The form of these questions is, "Which set of words is related to '*w*'?" The correct response was defined as the top three most similar words that were not morphologically

related, avoiding, for example, "infrequent" for the target word "frequent." Three sets of foils, or distractors, were chosen by selecting related words for other randomly chosen words of the same part of speech.

3. Evaluation

Two types of evaluations were conducted. First, the effectiveness of the thesaurus extraction technique was evaluated by comparing the output to a traditional thesaurus. Second, the quality of the automatically produced questions was measured in task-based evaluations involving teachers, native speakers, and non-native speakers of English.

Our evaluation of thesaurus extraction output follows the work of [7]. We used the electronically available Moby Thesaurus [16] as a gold standard, and then calculated precision values at ranks one through twenty. Precision is defined as the proportion of related words matching the gold standard out of the total number of related words produced by the thesaurus extraction system, so four words being correct at rank five would correspond to 80% precision. Specific to this evaluation is the definition of the gold standard. For a given word, Curran and Moens [7] used the union of the synonyms from three different thesauri in their evaluation. Such conflated synonym sets contained an average of over 300 synonyms for each word. This work uses a single thesaurus since using such large sets in the gold standard might produce overly optimistic results.

One hundred randomly selected words of various parts of speech were chosen for the evaluation sample. Mean and standard deviation values for precision at various ranks are shown in Table 2. The values are aggregated over the set words in the sample.

Table 2. *Mean and standard deviation of precision values of automatically extracted related words at various ranks.*

Rank (# related words)	Mean Precision	Std. Dev. Precision
1	0.43	0.50
3	0.29	0.18
5	0.25	0.26
10	0.18	0.56

The mean precision values are fairly low. Curran and Moens [7] reported values around 0.76 for precision at rank one. The most likely cause of this disparity is that the gold standard used in our evaluation had far fewer possible correct related words for each head word: 83 words on average versus over 300.

Another cause for the low mean precision values is that since the thesaurus extraction techniques do not exclusively extract synonyms, the output does not exactly match up with human-generated thesauri. The automatically extracted related words can also be antonyms, generalizations of the target word (i.e., hypernyms), and specifications (i.e., hyponyms). For example, for the head word "good," the automatic technique extracts the antonym "evil," which does not appear in the gold standard. Another example is that for the word "psychology", the automatic technique extracts other academic subjects such as "physics" and "chemistry". Although the extraction of non-synonymous related words is a disadvantage for the automatic technique in evaluations against traditional thesauri, this feature is actually useful for generating related word questions. It allows the system to generate sets of words that correspond to the second type of related words in Read's Word Associated Test [5]. Another advantage is that the possibility of multiple different relationships between the related words requires students to more deeply process the choices in the automatically generated related word questions.

However, the thesaurus extraction technique does make errors. Interestingly, it seems that the system is more prone to errors on some words than others, as the high standard deviation values in Table 1 indicate. One noticeable trend in our results is that the thesaurus extraction system appears to be more accurate at identifying certain parts of speech. We calculated the mean precision values of the top five extracted synonyms, as shown in Table 3. The system was most precise at identifying adjectives and adverbs (34.3%), and least accurate at identifying nouns (20.7%). One possible reason for this difference is that the set of possible related English nouns from which to choose related words is much larger than the set of verbs or adjectives.

Table 3. *Mean precision values of automatically extracted related words for different parts of speech.*

Part of Speech	Sample Size	Mean Precision at Rank 5
Adjective/Adverb	21	0.343
Noun	54	0.207
Verb	25	0.272

The system also performed relatively poorly for common polysemous verbs such as "put" or "take", which can mean different things depending on the use of particles, such as in the phrases "take out" or "take up". Also, using newswire texts for training data skewed the choice of some related words toward the business and political domains. For example, "investor" was chosen as one of the top related words for "partner".

After evaluating the thesaurus extraction methods, we evaluated the quality of the automatically generated questions produced by applying those methods. An example of an automatically generated question is shown in Figure 1. Questions were generated for words from the Academic Word List [17], a standard word list that is commonly used in ESL courses. First, a set of 50 questions was evaluated by an ESL instructor. He was told to indicate whether each question was of sufficient quality to be used for practice or assessment as part of the ESL course he teaches. He indicated that 68% of the questions were usable. He noted that, for most of the rejected questions, the problem was that the correct answer included one word that did not clearly relate to the target word. For example, the expected correct answer for "logic" was the set of words "reasoning", "wisdom", and "ethic", the third connection being unclear. Additionally, some questions involved errors that are easy to filter out, such as the inclusion of "import" in the answer set for "export," which would allow a student to guess the answer correctly using morphological rather than semantic cues.

Finally, the automatically generated questions were evaluated by natives and non-native speakers of English taking a short computerized test. The test included questions for twenty words randomly chosen from the Academic Word List [17]. It

contained sixty multiple choice questions, with three questions for each word. For each word, the student was asked a self-assessment question, a manually generated related word question, and an automatically generated related word question. The self-assessments asked, "Do you know the word '*w*'?" The formats of the two related word questions were identical to each other, and the ordering of the related word questions was randomized for each test-taker. Two native speakers took the test to ensure that the questions tested knowledge that native speakers have. Five non-native speakers also took the test. These subjects included two advanced/proficient non-natives as well as three upper-intermediate non-natives currently studying at the English Language Institute at the University of Pittsburgh.

The native speakers correctly answered 95% of the related word questions. Of the four items that were answered incorrectly, one was manually generated and three were automatically generated. One subject did not know one of the words on the test ("sector"), as indicated in the self-assessment portion of the test, and both questions were answered incorrectly for that word.

The non-native speakers, being fairly advanced, also performed well on the test. In self-assessments, they claimed to know 91% of the words. The non-native speakers correctly answered 85% of the manually generated questions, and 82% of the automatically generated questions. For a given word and student, the manually and automatically generated questions produced the same result, either correct or incorrect, more than 77% of the time. The agreement suggests that the automatically generated questions test roughly the same knowledge that the manually generated questions do. However, a more rigorous evaluation with a larger sample size and learners of different skill level is warranted.

4. Conclusions and Future Work

We have applied automatic thesaurus extraction techniques to automatically generate vocabulary questions that test knowledge of word associations. While there is room for improvement, the application shows promise as a means of producing practice and assessment materials without authoring by teachers. Native speakers are clearly able to identify related words, as shown by their very high performance in our evaluations. Non-natives also performed very well, likely due to their high proficiency levels. And we note that their performance was similar for manually and automatically generated questions.

Future work will be aimed at identifying areas for improvement for thesaurus extraction methods. We have noticed that the methods work surprisingly well for many words and often match human judgments, but they also perform poorly for many words. We plan to identify classes of words for which the extraction of related words performs poorly in order to improve the technique.

The use of statistical techniques such as automatic thesaurus extraction for generating assessments of word knowledge brings up the question of whether humans acquire certain features of vocabulary knowledge in a similar manner. The possibility of parallels between the processes of acquisition of lexical knowledge by humans and machines is of great interest and something we hope to explore further.

5. Acknowledgements

We would like to thank Gregory Mizera, Lois Wilson, and Alan Juffs from the English Language Institute for their assistance with the evaluations. This material is based on work supported by NSF grant SBE-0354420. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsor.

6. References

[1] Heilman, M., Collins-Thompson, K., Callan, J. & Eskenazi, M. (2006). Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension. *Proceedings of the Ninth International Conference on Spoken Language Processing*.

[2] Brown, J., Frishkoff, G., and Eskenazi, M. (2005). "Automatic question generation for vocabulary assessment." *Proceedings of HLT/EMNLP 2005*.

[3] Miller, G. (1990). "Wordnet: an on-line lexical database." *International Journal of Lexicography, 3(4)*.

[4] Liu, C., Wang, C., Gao, Z. and Huang, S. (2005). "Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items." *Proceedings of the Second Workshop on Building Educational Applications Using NLP*.

[5] Read, J. (1998). "Validating a test to measure depth of vocabulary knowledge." *Validation in language assessment*.

[6] Nation, P. (1990). *Teaching and learning vocabulary.* Rowley, MA: Newbury House.

[7] Curran, J. R., and Moens, M. (2002). "Improvements in Automatic Thesaurus Extraction." *Unsupervised Lexical Acquisition: Proceedings of the Workshop of the ACL Special Interest Groupon the Lexicon*.

[8] Harris, Z. S. (1968). *Mathematical Structures of Language*. Interscience Publishers.

[9] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1999). "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science*.

[10] Hindle, D. (1990). "Noun Classification from predicate-argument structures." *Proceedings of ACL-90*.

[11] Lin, D. (1998). "Automatic Retrieval and Clustering of Similar Words." *Proceedings of ACL-98.*

[12] Lin, D. (1993). "Principle-based parsing without overgeneralization." *Proceedings of ACL-93*.

[13] Graff, D. (2002). "English Gigaword." Linguistic Data Consortium. LDC Catalog No.: LDC2003T05.

[14] Pearce, D. (2002). A Comparative Evaluation of Collocation Extraction Techniques. Darren Pearce. *Third International Conference on Language Resources and Evaluation*.

[15] Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston, USA.

[16] Ward, G. (1996). *Moby Thesaurus*. Moby Project. http://www.dcs.shef.ac.uk/research/ilash/Moby/mthes.html

[17] Coxhead, A. (2000). "A New Academic Word List." *TESOL Quarterly*.