USING LEXICAL SEMANTIC TECHNIQUES TO CLASSIFY FREE-RESPONSES

JILL BURSTEIN
Educational Testing Service - llR
Princeton, New Jersey 08541

SUSANNE WOLFF
Educational Testing Service - 17R
Princeton, New Jersey 08541

AND

CHI LU
Educational Testing Service - 17R
Princeton, New Jersey 08541

Abstract. The education community wants to include more performance-based assessments on standardized exams. The research described in this paper shows the use of lexical semantic techniques for automated scoring of short-answer and essay responses from performance-based test items. We use lexical semantic techniques in order to identify the meaningful content of free-text responses for small data sets. The research demonstrates applications of lexical semantic techniques for free-text responses of varying length and in different subject domains. Prototype designs, and the results of the different prototype applications are discussed.

1. Introduction

Educational Testing Service (ETS) has begun to integrate more free-response (written responses) test questions on standardized exams. Due to the large volume of tests administered yearly by ETS, manual scoring of free responses is costly and time-consuming. ETS is interested in developing natural language understanding systems that could be used for computer-based scoring of free-responses (see Kaplan and Bennett, 1994; Burstein and Ka-

plan, 1995; Kud, 1995; and Burstein, *et al.*, 1998b).

This work falls into the general framework of information extraction in that we analyze unrestricted text in order to identify occurrences of specific types of content information as it is relevant to the domain of test questions.

The goal of the research was to develop an automated scoring system for short-answer free-response items. The paper describes a prototype system for classifying (i.e., scoring) short-answer responses, and its application to essay responses. The system needed to identify the relevant content of a response and assign it to an appropriate category. A driving consideration in the development of scoring system at the time of this study was that the available text corpora were relatively small. In addition, the responses in a given corpus typically exhibited a great deal of lexical and syntactic variation. The test items used in this paper were either computer-administered and experimental, or paper-and-pencil , "real" exams. In the former case, there was a limited subject pool, and in the latter case, we relied on a small data set of handwritten responses that had been manually transcribed into electronic form. The response sets available at the time of the study typically ranged from 300 - 700 responses. This is quite a different scenario from natural language understanding systems whose design can draw on data from large text corpora from sources such as the AP News and the Wall Street Journal.

More recently, we have had increasingly greater access to on-line free-text essay responses. Increased access to on-line data of free-text essay responses has given us a greater research capability. The essay responses that we are currently collecting have been evaluated in a more recently developed operational automated essay scoring system, the *e-rater system* described later in the paper.[1] *E-rater* was an outgrowth of the research in this paper. These initial studies taught us about the importance of the "domain-specificity" of test items. We also learned a considerable amount about how to design and develop systems for enhanced and reliable scoring of free-text, especially with regard to essay scoring.

[1]The *e-rater* system is a trademark of Educational Testing Service. The system may be referred to as *e-rater* or the *e-rater* system in this paper.

2. Test Item Types, Response Sets, and Lexical Semantics

2.1. TEST ITEM TYPES AND RESPONSE SETS

The experimental Formulating-Hypotheses (FH) item used for this study presents the examinee with a short text passage describing a situation, and prompts the examinee to provide up to 15 sentence-length responses to explain the situation. Examinees are supposed to rely on real-world knowledge about the item topic. Responses can be used to assess an examinee's ability to draw inferences. Both full-sentence and incomplete sentence responses were acceptable as long as the response content was appropriate. There is potentially no upper limit as to the number of plausible inferences that can be drawn, nor to the ways. These characteristics of FH item responses pose a unique challenge to the design of an automated scoring system.

We also applied the scoring techniques to essays responses from an Advanced Placement (AP) Biology exam. The exam requires students to write an essay on a topic in Biology – in this case, the topic focussed on *gel electrophoresis*. The test item specifies what points about the essay topic the student must discuss in the essay.

Language use may differ between test items and their associated response sets. Test items and their responses have been shown to be domain-specific (Kaplan and Bennett, 1994 and Burstein *et al*, 1997). However, the language in the small FH item response data sets is somewhat less predictable than a typical sublanguage with respect to vocabulary, collocations, and syntactic structures. At least with regard to the FH items (upon which these scoring techniques originated), vocabulary use and syntactic structures largely vary within the response set. There is less lexical and structural consistency in the responses. We therefore cannot take advantage of classification techniques that analyze language through the identification of similar patterns of vocabulary use, collocations, and syntactic structure. Sublanguage techniques such as Sager (1981) and Smadja (1993) used to represent the relevant information are therefore not applicable to our data sets.

2.2. USING LEXICAL SEMANTICS FOR RESPONSE REPRESENTATION

Broad coverage lexica that list the dictionary meaning of text words are not suitable for interpreting and representing the meaning of test responses. he application of a natural language processing application should not be restricted to a single domain. One wants to be able to re-use the application for any number of domains. In the context of the paper, "domains" refer

to the topics of test questions on standardized exams which span across any number of different topics. Gerstl gives two arguments with regard to why domain-specific strategies need to be applied to natural language processing systems. We believe that these apply to this work. First, with increasingly large data sets, there is an overwhelming amount of inconsistency across domains that becomes unmanageable. Domain-specific strategies contribute to the organization of the knowledge bases. Secondly, word meaning is heavily dependent on context (also, see Burstein and Kaplan (1995)). The domain in which a word occurs is part of the restriction of its semantic scope, or its conceptual representation given some domain (Cruse, 1986 and Gerstl, 1991).

The specialized lexical knowledge of test question response data sets is more precisely encoded by domain-specific concepts. These terms generalize over a set of text words in test item responses that are metonyms of each other. We define metonyms as words or multiword terms that can be substituted for each other in a given domain . In addition to being able to classify individual responses by content, an automated free-text response scoring system must be able to determine when responses are semantically identical (i.e., one response is the paraphrase of another response). A representation of responses by means of semantic scope or concepts greatly facilitates this task.

A specific example of the importance of metonym use can be explained using an example from responses to an Advanced Placement (AP) Biology essays question that we scored using the methods described in section 5.3. The AP Biology question required candidates to respond to a question about *gel electrophoresis*. It was a common and correct statement to explain in the response that the *DNA had been split into a certain number of fragments*. However, the term *fragments* was expressed in any number of ways, for example, *fragments, bands, segments, chains* and *pieces*. One can see how all these words might be used to refer to the word *fragments*. On the other hand, a general thesaurus is not as reliable a source from which the metonyms can be determined to be substitutable with *fragments*. We have found that domain-independent resources such as WordNet (Fellbaum, 1998), though very rich in information about word relationships, will not properly generate relevant metonym choices. We have found throughout our research in automated scoring of free-text that specialized lexical knowledge is most reliably found in the domain of the actual response data from a particular question.

We have also found that syntactic structure must be considered. In one

earlier system, concepts were represented without regard to syntactic structure. Also, the lexicon was domain independent. The underspecification of concept-structure relationships coupled with the lack of a domain-specific lexicon degraded the performance of that system (Kaplan and Bennett, 1994). Montemagni and Vanderwende (1993) have also pointed out that structural patterns are more desirable than string patterns for capturing semantic information from text. A lexically-based statistical approach which looked at similarity measures between responses based only on lexical overlap also performed poorly. Again, structure was not considered, and the lexicon was domain-independent which contributed to extremely low system performance (Burstein and Kaplan, 1995a and Burstein and Kaplan, 1995b). Clearly, the domain specific knowledge base has to be encoded in a specialized lexicon. The representation of response content is more reliably identified using concepts as they occur in syntactic structures.

Jackendoff's (1983) Lexical Conceptual Structure (LCS) representation would seem ideally suited for our task. LCSs are considered to be conceptual universals and have been successfully used by Dorr *et al.* (1995) and Holland (1994) in natural language understanding tasks. Holland points out though that LCSs cannot represent domain knowledge, nor can they handle the interpretation of negation and quantification, all of which are necessary in our scoring systems. Our system built to classify FH items handled negation and quantification in the concept grammar rules described in section 4.2. Holland also states that LCSs could not represent a near-match between the two sentences, *"The person bought a vehicle"*, and *"The man bought a car."* Since our scoring system must be able to deal with such near matches, and recognize possible metonyms, such as *vehicle* and *car*, these limitations of LCSs render Jackendoff's representation scheme unsuitable for our response classification problem.

3. The Formulating-Hypotheses Item

FH was an experimental inferencing item that consisted of a short passage (about 30 words) that described a hypothetical situation. [2] The examinee is asked to compose up to 15 hypotheses to explain why the situation might have occurred. Examinee responses did not have to be in the form of complete sentences. Responses could be up to 15 words in length. For example, the so-called police item describes a situation in which the number of police being killed had decreased over a 20-year period. The examinee was then

[2]Test items in this paper are copyrighted by Educational Testing Service (ETS). No further reproduction is permitted without written permission of ETS.

asked to give reasons as to why this might have occurred. Sample responses are given in (1).

(1) **Sample of correct responses to the police item**
   a. *Better cadet training programs.*
   b. *Police wear bullet-proof vests.*
   c. *Better economic circumstances mean less crime.*
   d. *Advanced medical technology has made it possible to save more lives.*
   e. *Crooks now have a decreased ability to purchase guns.*

Test developers [3] created a rubric (i.e., scoring key) which illustrated the criteria for classifying correct and incorrect responses into descriptive categories. They then manually classified each response according to the multi-category rubric. The scoring key was capable of capturing response duplication in the examinee's response set. For instance, if an examinee submitted the two responses, *Better trained police* and *Cops are more highly trained*, the scoring key must identify these two responses as paraphrases (duplicates), both of which should not count toward the final score. Multi-category scoring keys allow human graders to provide content-relevant explanations as to why a response was scored a certain way. The prototype system was designed to classify responses according to a set of training responses that had been manually scored by test developers in a multi-category scoring key. The system was required to handle duplicate responses. For the Police data set there were 32 categories associated with a set of 172 training responses.[4] Each scoring key category was represented by about 5 - 10 responses.

3.1. CHARACTERIZATION OF *POLICE* TRAINING DATA

The training set responses displayed a great deal of lexical and structural variation; therefore, co-occurrence patterns and frequencies did not yield consistent information about response content. For instance, *police* and *better* occur frequently, but in consistently varying syntactic structures such as *Police officers were better trained*, and *Police receiving better training to avoid getting killed in the line of duty.* According to the test developer's scoring key, these two responses were assigned to separate categories: (a) Better police training, general, and (b) Types of self-defense/safety techniques, respectively. Real world knowledge does not always suffice to establish metonymic relationships between terms. For instance, in the training

[3] Test developers write the test items and may also score them.

[4] There was a total of 47 rubric categories and 200 responses. We did not use 15 of the rubric categories since there were fewer than 5 responses per category. We also did not use the 28 responses from these 15 categories.

responses, *A recent push in safety training has paid off for modern day police* and *Officers now better combat trained*, the terms *safety training* and *combat trained* had been specified as "related." Test developers judged that *Safety training* and *combat train* both referred to a type of training for personal safety. They had categorized both responses under the *Trained for self-defense/safety category*. The terms had to be identified as metonyms in the response data in order to ensure correct classification of the responses.

4. Strategy for Representing *Police* Responses

A domain-specific concept lexicon be built based on the set of training responses. Every domain-relevant word or multiword term in the response set was linked to a concept entry in the lexicon. The semantically relevant content of responses is represented by a set of concept-structure rules in which lexical concepts (from the lexicon) are coupled with the syntactic structures in which they occur. A set of these rules forms a concept grammar. Small, individual concept grammars are developed for each of the scoring key categories.

The rules represent the syntactic relationship between concepts within and across phrasal constituents. Without this structural information, the concepts could occur in any position in a response, and automatic category assignment would not be reliable (Burstein and Kaplan, 1995). The automated procedure identifies conceptual and syntactic information, and retrieves concepts within specific phrasal and clausal categories, so that term clustering can be established.

4.1. THE SCORING LEXICON FOR THE POLICE ITEM

Our scoring lexicon is designed to capture Bergler's (1995) layered lexicon approach. The underlying idea in Bergler's approach is that the lexicon has several layers which are modular, and new layers can be plugged in for different texts. This allows lexical entries to be linked to text-specific information. While Bergler's domain is newspaper text, ours is free-response text for different test questions. Bergler points out that the regularities of of a task domain, such as newspaper text, need to be represented in terms of a sublanguage. In the case of assessment, we have found it to be the case that with each new test question comes a new micro-sublanguage in which words are defined within the domain of the test question topic. Free-responses to test questions have some characteristics of sublanguage, but at the same time are to less predictable with regard to syntactic structure and discourse style, than more formalized domains such as newpaper text.

In the layered lexicon approach, words are linked to definitions within some hierarchy. Bergler also has a meta-lexical layer which maps from syntactic patterns to semantic interpretation and does not affect the lexicon itself.

Our lexicon links words to superordinate concepts in the response set, as is shown below in the samples from the police item lexicon. A term denoting a concept is the head word of an entry. The concept is defined by a list of words and idioms that are metonyms in this domain. The metonyms that are subsumed under each concept are chosen from the entire set of individual words and idioms over the whole training set.

All words are listed in their base forms. Suffixation has been removed so that part of speech information did not interfere with conceptual generalizability. Since our sample was small and our domain highly specific to the domain of the test question, we found that it was necessary to buy as much conceptual generalizability as possible, perhaps at the expense of a small amount of word sense ambiguity. Word sense ambiguity did not seem to pose significant problems with these small response data sets.

In the spirit of the layered lexicon approach, the word list in the lexicon can remain constant and be reused. The links to concepts are modular, however, and can be changed given new domains. In our lexicon, terms denoting superordinate concepts are preceded by the pound sign (#). These concepts are followed by a list of subordinate metonyms terms. Head words not preceded by # are relevant words from the set of training responses which are subordinate metonyms of superordinate concepts. Head word entries will contain a pointer to a concept, indicated by the percent sign (%) preceding a superordinate concept name. Two sample entries are given in (2) below. The lexicon derived from the police training data contained fewer than 300 concepts.

(2) **Sample Entries from the *Police* Item Lexicon**

```
#BETTER [ better good advance improve effective increase
          efficient modern well increase... ]
ADVANCE [ %better ]
```

4.2. CONCEPT GRAMMAR RULES FOR THE POLICE ITEM

The concept grammar rule templates for mapping and classifying responses were built from 172 training set responses for 32 scoring key categories. The

training data was parsed using an industrial strength natural language processing tool (see MS-NLP (1996) for a description of this tool). Suffixes were removed by hand from the parsed data in this study. Based on the syntactic parses of these responses and the concept lexicon, we built a small concept grammar that characterizes responses by concepts and relevant structural information for each scoring key category. The phrasal constituents were identified by the general XP label. Sample concept grammar rules are shown in (3).

(3) **Sample Concept Grammar Rules for** *Types of self-defense/safety*

a. XP:[POLICE],XP: [BETTER,TRAIN],XP: [SAFETY]
b. XP:[TRAIN],XP:[POLICE,SAFETY], XP: [BETTER, SAFETY]
c. XP:[POLICE,BETTER,TRAIN], XP: [SAFETY,DANGER,SITUATION]
d. XP:[SPECIALIST],XP:[TRAIN,SAFETY]

Our concept grammars could be compared to Bergler's meta-lexical layer (see section 4.1) in that they provide a mapping between the syntax and the semantics of responses.

4.3. PROCESSING RESPONSES FOR CATEGORY ASSIGNMENT

The program extracts words and terms in Noun Phrases (NP), Verb Phrases (VP), Prepositional Phrases (PP), Infinitive Clauses (INFCL), Subordinate Clauses (SUBCL), Adjective Phrases (ADJP) and Adverb Phrases (ADVP). All phrasal and clausal constituent nodes are then collapsed into a generalized representation, XP. All single XPs and combinations of multiple XPs were matched against the concept grammars for each content category to locate rule matches. This procedure is illustratedin (4) below.

The type of phrasal constituent, that is, whether it is an NP or a VP, was not informative, given these small data sets. In fact, when we left in information about syntactic category, performance was lower. Fewer matches occured between concept grammar rules with specific constituent information, than occurred when the constituent information was collapsed into the general XP category.

(4) **Scoring Procedure**

a. Input Response:
   (a) *Cops are better trained in self-defense*
b. Tagged Phrasal Nodes of Parsed Response:

[Cops=POLICE]NP

[better=BETTER,trained=TRAIN]VP

[self-defense=SAFETY]PP

c. Collapsed Phrasal Nodes:

(a) XP: [Cops=POLICE]

(b) XP: [better-BETTER,trained=TRAIN]

(c) XP: [self-defense=SAFETY]

d. Matched Tagged Nodes to Concept Grammar Rules (see (3), above):

(a) XP: [POLICE], XP:[BETTER,TRAIN],XP:[SAFETY]

4.4. DOES MANUAL PREPROCESSING OUTWEIGH THE BENEFITS OF AUTOMATIC SCORING?

The automated scoring procedure required response data to be pre-processed by hand. This raises the issue of cost and time efficiency. The total person-time and computing time must be considered in relation to how long it would take test developers to manually classify a data set. Two people completed the manual creation of the lexicon and the concept grammar rules for this data set in approximately 40 hours – about 20 hours per person. After this initial study, we developed a program to automate the generation of the concept grammar rules described in section 5.3. Previously, concept grammar rules had been manually created. This program to generate the concept grammar rules cut the pre-processing time in half. For a similar test item, it would take one person approximately 8 -10 hours to create the lexicon, and another 8-10 hours to do the pre-processing (that is, do some manual revisions to the automatically created concept grammar rules).

The FH item was an experimental pilot item for the Graduate Record Examination (GRE). If such an item were to become an actual test item, we would envision the following scenario with regard to automated scoring. The GRE is administered to approximately 28,000 examinees per year. Since every examinee can give up to 15 responses for an FH item, approximately 420,000 responses for this item could be collected over the year. Each examinee's response set would then typically be scored by two human graders. It is difficult to estimate how long the manual scoring process would take in hours, but it is safe to assume that it would considerably exceed 20 hours allocated for manual processing by our prototype that includes a procedure for automatic concept grammar generation. Automated scoring appears to be a viable cost-saving and time-saving option, even if

some manual pre-processing is required. Practically speaking, , for operational scoring, it is highly desirable that all pre- and post-processing scoring system components are fully automated.

5. Results

5.1. POLICE ITEM: INITIAL RESULTS

There were 172 training responses for the *police* item. The training set provided the vocabulary for building the concept lexicon and the concept grammar rules. The test data consisted of an independent set of 206 test responses from 32 content categories. The results of the automated scoring are presented in Table 1.

Response	Set Coverage	Accuracy
Training and Test Set	92%(347/378)	90%(313/347)
Test Set	87%(180/206)	81%(146/180)

TABLE 1. Results of automatic scoring of *police* responses

5.1.1. *Analysis*

We have assessed that errors that degrade the system performance are due to either a) lexical gaps, b) human grader misclassification, c) concept-structure problems, or (d) cross-classification.

A lexical gap error characterizes cases where a response could not be classified because it was missing a concept tag,. Therefore the responses did not match a rule in the grammar. Forty percent of overall errors were lexical gap errors. Since the lexicon is built from the vocabulary in the training set, it is not surprising to find that the words not recognized by the system occurred only in the test data sets. These metonyms were not identified as having synonymous relations in any of our on-line thesaurus or dictionary sources, such as WordNet. For instance, the response *Police are better skilled* was not scored because the phrase *better skilled* did not occur in the training set. Consequently, *skill* was not listed in the lexicon as a metonym of *train*, and the response could not be recognized as a paraphrase of *better trained police*. It is expected that results would improve if concepts were represented by a larger number of metonyms. This prediction was proven when we augmented the lexicon with metonyms that

could be accessed from the test data. Rerunning the prototype system after the addition of just 56 metonyms yielded improved results (see Table 2).

Concept structure rule problems made up 30% of the errors. Such problems occurred when a response could not be classified because its concept-structure patterning did not match any of the existing concept-structure rules. Significant conceptual similarity between two scoring key categories and the potential for categorical cross-classification accounted for another 17% of the errors. In one percent of the cases, the human graders had misclassified a response in the training set and also the test set response was misclassified. For instance, the response *Officers are better trained and more experienced so they can avoid dangerous situations* was misclassified under the scoring key category *Better trained police, general* instead of the category *Better intervention/crook counseling.*

5.1.2. *Results Using an Augmented Lexicon*
As seen above, 40% of the errors could be accounted for by lexical gaps. We expected our results to improve if more metonyms of existing concepts were added to the lexicon. Therefore, we augmented the lexicon with metonyms from the test data. We re-ran the scoring program, using the augmented lexicon on the same set of data. The results of this run are presented in Table 2.

Response set	Coverage	Accuracy
Training and Test Set	96%(364/378)	96%(341/364)
Test Set	93%(193/206)	83%(178/193)

TABLE 2. Automated scoring results using an augmented lexicon

We also used the augmented lexicon from this second experiment to score a set of test data that had not been classified by test developers. This was the only additional data available to test the generalizability of the methods used for automated scoring. This experiment would not allow us to measure agreement with human scoring decisions. Coverage for responses given a single classification by the procedure or multiple classifications is 70% and 78% respectively. Accuracy in Table 3 indicates the number of classifications that were judged informally by the authors as "relevant" with regard to the individual categories established by the test developers.

	Coverage	Accuracy
Single Classifications	70%(303/436)	75%(228/303)
Total Classifications	78%(340/436)	77%(262/340)

TABLE 3. Results of unscored response sets

Recall, that prototype scoring system was trained on a small data set consisting of 172 responses. Given the small data set, it may not be able to recognize concepts and concept structure patterns due to lexical gaps or missing concept grammar rules.

5.2. FURTHER EXPLORATIONS

Subsequently, we applied this technique to two additional free-response data sets. One data set was a set of FH responses from a different FH question. The second set were essay responses from an Advanced Placement essay exam.

5.2.1. *The Artist Item*

The *artist item* is another FH item on which we tested our automated scoring methods. The total response set of 428 responses was partitioned into a 200-response training set and a 228 test set. The methods employed were identical to those employed for the *police* data in that we derived the scoring lexicon for this data from a training corpus of responses. We manually wrote the concept structure rules for each of the 43 scoring key categories established by human graders in another few days. The rules were done manually since an automatic rule generation program had not yet been implemented. All other components of the prototype system remained unchanged. The results are presented in Table 4.

	Coverage	Accuracy
Training Set	94%(187/200)	94%(176/187)
Test Set	56%(128/228)	48%(61/128)

TABLE 4. Automated scoring results artist data

5.2.2. *Discussion*

Creation of a lexicon and concept grammar based on the training set of 200 sentences took about 40 hours. The other components of our automated scoring system did not require customization. On the other hand, the conceptual difference between some categories was minute, perhaps more so than for the police item. The scoring system has to rely on information that is lexically and structurally available in the text. It cannot draw on real world knowledge. So while a human could infer that the response *A great deal of copies* belonged to the scoring key category *Dealers, others, faked documents, works*, the machine could not. Lexical gaps, gaps in the concept structure grammar, and misclassifications by test developers are again responsible for degraded results achieved in the automatic scoring of the blind data set.

5.3. ESSAY RESPONSE DATA

The lexical semantic techniques for sentence length responses described above were also applied to score essay responses for a College Board Advanced Placement (AP) Biology test item. A detailed discussion of this study can be found in Burstein *et al.* (1997). This item was a suitable candidate for automated scoring using the techniques applied to the FH responses, especially since there was an increased amount of lexico-syntactic patterning that could be identified amongst these responses using a concept lexicon.

The item had been administered as a paper-and-pencil item and had to be manually transcribed into electronic files. The length of the essays and the transcription effort restricted the pool of essays available for our experiment to 200 that had been rated "Excellent" by human graders. The essays were divided into a training and a test set. For comparison, we added an available small set of "Poor" essays.

The test item was subdivided into 4 sections, each of which corresponded with a prompt that asked the examinee to explain or describe certain aspects of *gel electrophoresis.* The human reader scoring guide followed the organization of the test item. Examinees typically divided the responses into the sections corresponding to the discussion points as they were organized in the test item. This allowed us to partition the electronic files into sections for automated scoring.

We developed a lexicon based on the training data essay responses. We generated the concept-structure rules composing the concept-structure

grammars to represent a computer-based scoring key corresponding to the human reader scoring key. Analysis of the *police* and *artists* data had shown that in creating the concept grammars by hand, as we had done initially, we had occasionally omitted a concept structure rule that was a permutation of an existing rule. Therefore, we implemented an automated procedure for generating all permutations of concept structure rules. This was significantly faster and also more accurate than manual creation. Overgeneration of rules turned out to be unproblematic. The results for 85 "Excellent" test essays and 20 "Poor" tests essays are shown in Table 5. Accuracy measures the exact agreement between machine and human scores. The columns labeled "Accuracy +/-1" and "Accuracy +/-2" show agreement between machine and human scorers within 1 or 2 points of the human score, respectively.

Test Set	Coverage	Accuracy	Accuracy +/-1	Accuracy +/-2
Excellent	100%	89%(76/85)	95%(81/95)	100%(85/85)
Poor	100%	75%(15/20)	90% (17/20)	95%(18/20)
Total	100%	87%(91/105)	94%(99/105)	96%(103/105)

TABLE 5. AP Biology scoring results for the test data sets

5.3.1. *Results*

The very nature of this test question with its conceptually specific scoring key categories made it a very good candidate for these automated scoring procedures. In addition, the repeated patterns of lexico-syntactic information in responses was highly compatible with the technique used to score FH responses. Accordingly, the prototype scoring system used for AP Biology essays shows very high agreement with human rater scores. Agreement between machine scores and human scores for adjacent scores was considerably higher. As with the scoring of the police data, lexical gaps and concept grammar rule deficiencies were primarily responsible for scoring errors. However, the automatic rule generation used in the AP Biology study appeared to greatly enhance the concept structure grammars, and performance was increased.

6. Conclusion

Our initial results for scoring the *police* data are encouraging and lend support to the hypothesis that lexical semantic techniques can be inte-

grated into an automated system for scoring free-response test questions of the type described in this paper. Our system for scoring short-answer free responses makes use of concepts and concept structure patterns to identify relevant meaning elements of a response. Concepts represent a level of abstraction over text words and expressions, since in a given domain they function as metonyms. Concepts and metonyms are listed in a domain specific lexicon. We also need to experiment with lexical techniques that could help us automate the development of item specific lexicons and the extension of concepts in these lexicons, so that we can overcome the lexical gap problem. Perhaps lexical gap errors could be reduced by using example-based methods (Richardson *et al.*, 1993; and, Tsutsumi, 1992) or corpus-based techniques (Church and Hanks, 1990) to build the concept lexicons.

The artists data results were poor, perhaps due to a combination of data sparseness, deficient lexical entries and excessive numbers of classification categories. However, our experiments with the FH police item and with the AP Biology item show that the automated scoring system can perform well for diverse test items and subject domains, given sufficient data and reasonable classification categories. To the extent that we can automate lexicon creation and concept structure rule generation and induction (Soderland *et al.*, 1994) along the paths indicated, it could prove to be a useful tool for scoring free-response test items.

Most recently, we have developed the *e-rater* system, an automated essay scoring system being used to score essays on the Graduate Management Admissions Test (GMAT) (Burstein, *et al.*, 1998a; Burstein, *et al.*, 1998b; and, Burstein, *et al.* 1998c). In initial studies using 9573 essays from 15 different essay topics from native and nonnative English speakers, *e-rater*'s exact plus adjacent agreement with human scores ranges between 87% and 94%. We achieved increased performance recently in the operational version of the system and have found that the *e-rater's* exact and adjacent agreement with human reader scores is as high as 95by exact plus adjacent agreement as well.

This is a very different method of analyzing free-text responses than was used in this study. It is completely automated. The system is trained on human reader scored essays for every essay question. Scores are based on the system's evaluation of syntactic, discourse, and topical analysis features. The system also uses a lexicon of cue words that are classified by their discourse functions. The lexical classifiers characterize discourse relationships suggested in (Quirk and Greenbaum, 1985).

To access the topical content of an essay, the *e-rater* system uses content vector analysis methods at the level of the whole essay, as well as at the level of individual arguments in the essay in order to identify relationships between essay score and word use in a domain (test question topic). The technique used in *e-rater* for topical analysis by argument links high level discourse structure of the essay with essay vocabulary. The current version of the the topical analysis componenet does not have a mechanism to recognize metonyms within the domain of the test question. Other systems that evaluate word use across text and documents use techniques such as Singular Value Decomposition (SVD) to evaluate synonym relationships between words across texts. These methods have also been applied to essay scoring (Foltz, *et al*, 1998). We are currently experimenting with methods that might be used to enhance the *e-rater* system, so that the topical analysis component of the system can identify metonyms using a general thesaurus of essay responses for test questions. These techniques might enable the system to automatically locate metonyms for each domain of a new essay question. We anticipate that this kind enhancement for metonym recognition might increase the performance of the the topical analysis component of the *e-rater* system. An enhanced topical analysis component that could detect metonyms in different test question domains would yield more precise information about domain-specific word use, as we found was useful in the FH item and AP essay studies described in this paper. Such information could be helpful to generate diagnostic and instructional feedback about essays.

References

Bergler, Sabine. (1995). From Lexical Semantics to Text Analysis. In Patrick Saint-Dizier and Evelyne Viegas (eds.), *Computational Lexical Semantics*, Cambridge University Press, New York, NY.

Burstein, Jill, Karen Kukich, Lisa Braden-Harder, Martin Chodorow, Shuyi Hua, Bruce Kaplan, Chi Lu, James Nolan, Don Rock and Susanne Wolff. (1998a). Computer Analysis of Essay Content for Automated Score Prediction: A prototype automated scoring system for GMAT Analytical Writing Assessment. (RR-98-15). Princeton, NJ: Educational Testing Service.

Burstein, Jill, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris (1998b). Automated Scoring Using A Hybrid Feature Identification Technique. In the Proceedings of *36th the Annual Meeting of the Association of Computational Linguistics*, Montréal, Canada.

Burstein, Jill, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow (1998c). Enriching Automated Scoring Using Discourse Marking. In the Proceedings of the *Workshop on Discourse Relations Discourse Marking*, 36th Annual Meeting of the Association of Computational Linguistics, Montréal, Canada.

Burstein, Jill C., Susanne Wolff, Chi Lu, Randy M. Kaplan. (1997). An Automatic Scoring

System for Advanced Placement Biology Essays. In the proceedings of the *Fifth Conference on Applied Natural Language Processing*, Washington, DC.

Burstein, Jill C. and Randy M. Kaplan. (1995a). On the Application of Context to Natural Language Processing Applied to the Analysis of Test Responses. In Proceedings from the *Workshop on Context in Natural Language Processing*, IJCAI, Montréal, Canada.

Burstein, Jill C. and Randy M. Kaplan. (1995b). GE-FRST Evaluation Report: How well does a Statistically-Based Natural Language Processing System Score Natural Language Constructed Responses? (RR-95-29). Princeton, NJ: Educational Testing Service.

Church, K. and P. Hanks. (1990). Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 16(1), 22-29.

Cruse, D.A. (1986). *Lexical Semantics.* Cambridge University Press, Cambridge, UK.

Fellbaum, Christiane. (1998). : An Electronic Lexical Database. MIT Press, Cambridge, MA.

Foltz, P. W., W. Kintsch, and T. K. Landauer. (1998). The Measurment of Textual Coherence with Latent Semantic Analyis. *Discourse Processes*, 25(23), 285-307.

Gerstl, P. (1991). A Model for the Interaction of Lexical and Non-Lexical Knowledge in the Determination of Word Meaning. In J. Pustejovsky and S. Bergler (Eds), *Lexical Semantics and Knowledge Representation*, Springer-Verlag, New York, NY.

Holland, Melissa V. (1994) Intellignet Tutors for Foreign Languages: How Parsers and Lexical Semantics can Help Learners and Assess Learning. In Proceedings of *the Educational Testing Service Conference on Natural Language Processing Techniques and Technology in Education and Assessment*, Princeton, New Jersey.

Kaplan, Randy M. and Randy E. Bennett. (1994). Using the Free-Response Scoring Tool To Automatically Score the Formulating-Hypothesis Item. (RR-94-08). Princeton, NJ: Educational Testing Service.

Kud, Jacquelynne M., George G. Kripka, and Lisa Rau. (1994). Methods Short Answer Responses for Categorizing. In Proceedings of *the Educational Testing Service Conference on Natural Language Processing Techniques and Technology in Education and Assessment*, Princeton, New Jersey.

Montemagni, Simonetta and Lucy Vanderwende. (1993). Structural Patterns versus String Patterns for Extracting Semantic Information from Dictionaries, in K. Jensen, G. Heidorn and S. Richardson (eds) *Natural Language Processing: The PLNLP Approach*, Kluwer Academic Publishers, Boston, MA.

MSNLP (1996): http ://research.microsoft.com/research/nlp

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartik (1985). *A Comprehensive Grammar of the English Language.* Longman, New York.

Richardson, Stephen D., Lucy Vanderwende, and William Dolan. (1993). Combining Dictionary bases and Example-Based Methods for Natural Language Analysis. (MSR-TR-93-08), Redmond, WA, Microsoft Corporation.

Sager, N. (1981). *Natural Language Information Processing: A computer grammar of English and its applications.* Addison-Wesley, Reading, Massachusetts.

Smadja, Frank. (1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics* 19(1), 143-177.

Soderland, Stephen, David Fisher, Jonathan Aseltine and Wendy Lehnert. (1994). CRYSTAL: Inducing a Conceptual Dictionary, available from http://ciir.cs.umass.edu/medical/ss-ijkai.hmtl

Tsutsumi, T. (1992). Word Sense Disambiguation by Examples, in K. Jensen, G. Heidorn and S. Richardson (eds), *Natural Language Processing: the PLNLP Approach*, Kluwer Academic Publishers, Boston, MA.