An Automatic Scoring System For Advanced Placement Biology Essays

Jill Burstein, Susanne Wolff, Chi Lu
Educational Testing Service MS-11R
Princeton, NJ 08541
e-mail: jburstein@ets.org

Randy M. Kaplan
Advanced Technology
PECO Energy
Philadelphia, PA

Abstract

This paper describes a prototype for automatically scoring College Board Advanced Placement (AP) Biology essays.[1]. The scoring technique used in this study was based on a previous method used to score sentence-length responses (Burstein, et al, 1996). One hundred training essays were used to build an example-based lexicon and concept grammars. The prototype accesses information from the lexicon and concept grammars to score essays by assigning a classification of Excellent or Poor based on the number of points assigned during scoring. Final computer-based essay scores are based on the system's recognition of conceptual information in the essays. Conceptual analysis in essays is essential to provide a classification based on the essay content. In addition, computer-generated information about essay content can be used to produce diagnostic feedback. The set of essays used in this study had been scored by human raters. The results reported in the paper show 94% agreement on exact or adjacent scores between human rater scores and computer-based scores for 105 test essays. The methods underlying this application could be used in a number of applications involving rapid semantic analysis of textual materials, especially with regard to scientific or other technical text.

INTRODUCTION

To replace the conventional multiple choice questions on standardized examinations, Educational Testing Service (ETS) is currently developing computer-based scoring tools for automatic scoring of natural language constructed-responses -- responses that are written, such as a short-answer or an essay. The purpose of this work is to develop computer-based methods for scoring so that computer-administered natural language constructed-response items can be used on standardized tests and scored efficiently with regard to time and cost.

Until recently, ETS's automated scoring efforts were primarily devoted to the development of computer programs used to score short-answer constructed-responses of up to 15 words (Burstein and Kaplan, 1995 and Burstein et al., 1996). In this study a classification of *Excellent* or *Poor* was automatically assigned to an AP Biology essay. Our initial goal in this study was to develop a prototype scoring system that could reliably assign a classification of Excellent to a set of AP Biology essays. For the evaluation of the scoring method, a small sample of Poor essays were also scored to compare the results.[2]

Human rater scoring of AP Biology essays is based on a highly constrained scoring key, called a rubric, that specifies the criteria human raters use to assign scores to essays. Accordingly, for the test question studied here, the criteria for point

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

[2] The Poor classification is not an official AP classification. It was used in this study to distinguish the Excellent essays with scores of 9 and 10 from essays with lower end scores in the 0 - 3 range.

assignment are highly constrained. Essentially, the essay can be treated as a sequence of short-answer responses. Given our preliminary successes with test questions that elicit multiple responses from examinees, similar scoring methods were applied for scoring AP Biology essay. The results show 87% agreement for exact scores between human rater and computer scores, and 94% agreement for exact or adjacent scores between human rater and computer scores.

This work is also applicable for other types of assessment as well, such as for employee training courses in corporate and government settings. Since the methods discussed in this paper describe techniques for analysis of semantic information in text, presumably this application could be extended to public informational settings, in which people might key in "requests for information" in a number of domains. In particular, these methods could be successfully applied to the analysis of natural language responses for highly constrained domains, such as exist in scientific or technical fields.

SYSTEM TRAINING

One hundred Excellent essays from the original 200 essays were selected to train the scoring system. The original 200 essays were divided into a training set and test set, selected arbitrarily from the lowest examinee identification number. Only 85 of the original 100 in the test set were included in the study due to illegibility, or use of diagrams instead of text to respond to the question. For convenience during training, and later, for scoring, essays were divided up by section, as specified in the scoring guide (see Figure 1), and stored in directories by essay section. Specifically, the Part A's of the essays were stored in a separate directory, as were Part B's, and Part C's. Examinees typically partitioned the essay into sections that corresponded to the scoring guide.

System training involved the following steps that are discussed in subsequent sections: a) manual lexicon development, b) automatic generation of concept-structure representation (CSR), c) manual creation of a computer-based rubric, d) manual CSR "fine-tuning", e) automatic rule generation, and f) evaluation of training process.

Lexicon Development

Example-based approaches to lexicon development have been shown to effectively exemplify word meaning within a domain (Richardson, et al., 1993, and Tsutsumi 1992). It has been further pointed out by Wilks, et al, 1992, that word senses can be effectively captured on the basis of textual material. The lexicon developed for this study used an example-based approach to compile a list of lexical items that characterized the content vocabulary used in the domain of the test question (i.e., gel electrophoresis). The lexicon is composed of words and terms from the relevant vocabulary of the essays used for training.

To build the lexicon, all words and terms considered to contribute to the core meaning of each relevant sentence in an essay, were included in the lexicon. The decision with regard to whether or not a sentence was relevant was based on information provided in the scoring guide (in Figure 1). For instance, in the sentence, "*Smaller DNA fragments move faster than larger ones.*", the terms *Smaller, DNA, fragments, move, faster, larger* are considered to be the most meaningful terms in the sentence. This is based on the criteria for a correct response for the Rate/Size category in the scoring guide.

Each lexical entry contained a superordinate concept and an associated list of metonyms. *Metonyms* are words or terms which are acceptable substitutions for a given word or term (Gerstl, 1991). Metonyms for concepts in the domain of this test question were selected from the example responses in the training data This paradigm was used to identify word similarity in the domain of the essays. For instance, the scoring program needed to recognize that sentences, such as *Smaller DNA fragments move faster than larger ones* and *The smaller segments of DNA will travel more quickly than the bigger ones,* contain alternate words with similar meanings in the test question domain. To determine alternate words with similar meanings, metonyms for words, such as *fragments* and *move* were established in the

lexicon so that the system could identify which words had similar meanings in the test item domain. The example lexical entries in (1) illustrate that the words *fragment* and *segment* are metonyms in this domain, as well as the words *move* and *travel*. In (1), FRAGMENT and MOVE are the higher level lexical concepts. The associated metonyms for FRAGMENT and MOVE are in adjacent lists illustrated in (1).

(1). Sample Lexical Entries

FRAGMENT [*fragment* particle *segment* ...]

MOVE [*move travel* pass pull repel attract ...]

Concept-Structure Representations (CSR)

Obviously, no two essays will be identical, and it is unlikely that two sentences in two different essays will be worded exactly alike. Therefore, scoring systems must be able to recognize paraphrased information in sentences across essay responses.. To identify paraphrased information in sentences, the scoring system must be able to identify similar words in consistent syntactic patterns. As, Montemagni and Vanderwende (1993) have also pointed out, structural patterns are more desirable than string patterns for capturing semantic information from text. We have implemented a concept-extraction program for preprocessing of essay data that outputs conceptual information as it exists in the structure of a sentence. The program reads in a parse tree generated by Microsoft's Natural Language Processing Tools (MSNLP) for each sentence in an essay.[3] The program substitutes words in the parse tree with superordinate concepts from the lexicon, and extracts the phrasal nodes containing these concepts. (Words in the phrasal node which do not match a lexical concept are not included in the set of extracted phrasal nodes.) The resulting structures are CSRs. Each CSR represents a sentence according to conceptual content and phrasal constituent structure. CSRs characterize paraphrased information in sentences. For example, in the sentences "*The DNA segment would be digested only once, leaving 2 pieces.*", and "*The DNA fragment would only have 2 segments,*" the phrases *DNA segment* and *DNA fragment* are paraphrases of each other, and *2 pieces* and *2 segments* are paraphrases of each other. These sentences are represented by the CSR in (2a) and in (2b).

(2)a. NP:[DNA,FRAGMENT]
NP:[TWO,FRAGMENT]

In the final version of the CSR, phrasal constituents are reduced to a general XP node, as is illustrated in

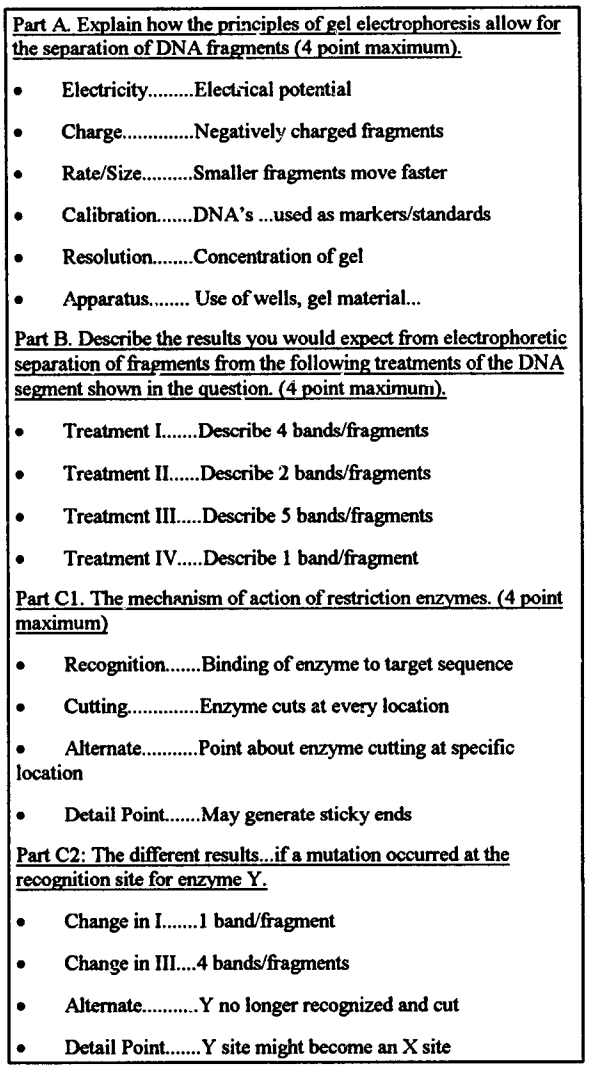
(2)b..XP:[DNA,FRAGMENT]
XP:[TWO,FRAGMENT]

Since phrasal category does not have to be specified, the use of a generalized XP node minimizes the number of required lexical entries, as well as the number of concept grammar rules needed for the scoring process.

The Computer Rubric

Recall that a rubric is a scoring key. Rubric categories are the criteria that determine a correct response. A computer-based rubric was manually created for the purpose of classifying sentences in essays by rubric category during the automated scoring process. Computer rubric categories are created for the bulleted categories listed in the human rater scoring guide illustrated in Figure 1.

[3] See http://research.microsoft.com/research/nlp for information on MS-NLP.



Part A. Explain how the principles of gel electrophoresis allow for the separation of DNA fragments (4 point maximum).

- Electricity.....Electrical potential
- Charge.....Negatively charged fragments
- Rate/Size.....Smaller fragments move faster
- Calibration.....DNA's ...used as markers/standards
- Resolution.....Concentration of gel
- Apparatus..... Use of wells, gel material...

Part B. Describe the results you would expect from electrophoretic separation of fragments from the following treatments of the DNA segment shown in the question. (4 point maximum).

- Treatment I.....Describe 4 bands/fragments
- Treatment II.....Describe 2 bands/fragments
- Treatment III.....Describe 5 bands/fragments
- Treatment IV.....Describe 1 band/fragment

Part C1. The mechanism of action of restriction enzymes. (4 point maximum)

- Recognition.....Binding of enzyme to target sequence
- Cutting.....Enzyme cuts at every location
- Alternate.....Point about enzyme cutting at specific location
- Detail Point.....May generate sticky ends

Part C2: The different results...if a mutation occurred at the recognition site for enzyme Y.

- Change in I.....1 band/fragment
- Change in III....4 bands/fragments
- Alternate.....Y no longer recognized and cut
- Detail Point.....Y site might become an X site

Figure 1: Scoring Guide Excerpt

Accordingly, the computer-rubric categories were the following. For Part A, the categories were *Electricity, Charge, Rate/size,Calibration, Resolution, and Apparatus.* For Part B the categories were, *Treatment I, Treatment 2, Treatment 3, and Treatment IV.* For Part C1, the categories were: *Recognition, Cutting, Alternate, and Detail Point.* For Part C2, the categories were *Change in I, Change in II, Alternate, and Detail Point.* Each computer-rubric category exists as an electronic file and contains the related concept grammar rules used during the scoring process. The concept grammar rules are described later in the paper.

Fine-Tuning CSRs

CSRs were generated for all sentences in an essay. During training, the CSRs of relevant sentences from the training set were placed into computer-rubric category files. Relevant sentences in essays were sentences identified in the scoring guide as containing information relevant to a rubric category. For example, the representation for the sentence, "*The DNA fragment would only have 2 segments,*" was placed in the computer rubric category file for Treatment II.

Typically, CSRs are generated with extraneous concepts that do not contribute to the core meaning of the response. For the purpose of concept grammar rule generation, each CSR from the training data must contain only concepts which denote the core meaning of the sentence. Extraneous concepts had to be removed before the rule generation process, so that the concept-structure information in the concept grammar rules would be precise.

The process of removing extraneous concepts from the CSRs is currently done manually. For this study, all concepts in the CSR that were considered to be extraneous to the core meaning of the sentence were removed by hand. For example, in the sentence, *The DNA segment would be digested only once, leaving 2 pieces,* the CSR in (3) was generated. For *Treatment II,* the scoring guide indicates that if the sentence makes a reference to *2 fragments* that it should receive one point. (The word, *piece,* is a metonym for the concept, *fragment,* so these two words may be used interchangably.) The CSR in (3) was generated by the concept-extraction program. The CSR in (4) (in which XP:[DNA,FRAGMENT] was removed) illustrates the fine-tuned version of the CSR in (3). The CSR in (4) was then used for the rule generation process, described in the next section.

(3) XP:[DNA,FRAGMENT]
XP:[TWO,FRAGMENT]

(4) XP:[TWO,FRAGMENT]

Concept Grammar Rule Generation

At this point in the process, each computer rubric category is an electronic file which contains fine-tuned, CSRs. The CSRs in the computer rubric categories exemplify the information required to receive credit for a sentence in a response. We have developed a program that automatically generates rules from CSRs by generating permutations of each CSR The example rules in (5) were generated from the CSR in (4). The rules in (5) were used during automated scoring (described in the following section).

(5)a. XP:[TWO, FRAGMENT]

b. XP:[FRAGMENT,TWO]

The trade-off for generating rules automatically in this manner is rule overgeneration, but this does not appear to be problematic for the automated scoring process. Automated rule generation is significantly faster and more accurate than writing the rules by hand. We estimate that it would have taken two people about two weeks of full-time work to manually create the rules. Inevitably, there would have been typographical errors and other kinds of "human error". It takes approximately 3 minutes to automatically generate the rules.

AUTOMATED SCORING

The 85 remaining *Excellent* test essays and a set of 20 *Poor* essays used in this study were scored. First, all sentences in Parts A, B and C of each essay were *parsed* using MSNLP. Next, inflectional suffixes were automatically removed from the words in the parsed sentences, since inflectional suffixed forms are not included in the lexicon. CSRs were automatically generated for all sentences in each essay. For each part of the essay, the scoring program uses a searching algorithm which looks for matches between CSRs and/or subsets of CSRs, and concept grammar rules in rubric categories associated with each essay part. Recall that CSRs often have extraneous concepts that do not contribute to the core meaning of the sentence. Therefore, the scoring program looks for matches between concept grammar rules and subsets of CSRs, if no direct match can be found for the complete set of concepts in a CSR. The scoring program assigns points to an essay as rule matches are found, according to the scoring guide (see Figure 1). A total number of points is assigned to the essay after the program has looked at all sentences in an essay. Essays receiving a total of at least 9 points are classified as Excellent, essays with 3 points or less are classified as Poor, and essays with 4 - 8 points are classified as "Not Excellent." The example output in Appendix 1 illustrates matches found between sentences in the essay and the rubric rules from an Excellent essay.

RESULTS

Table 1 shows the results of using the automatic scoring prototype to score 85 *Excellent* test essays, and 20 *Poor* test essays. Coverage (Cov) illustrates how many essays were assigned a score. Accuracy (Acc) indicates percentage of agreement between the computer-based score and the human rater score. Accuracy within 1 (w/i 1) or 2 points (w/i 2) shows the amount of agreement between the computer scores and human raters scores, within 1 or 2 points of human rater scores, respectively. For Excellent essays computer-based scores would be 1 or 2 points below the 9 point minimum, and for Poor essays, they would be 1 or 2 points above the 3 point maximum.

Data Set	Cov	Acc	Acc w/i 1	Acc w/i 2
Excellent	100%	89%	95%	100%
Poor	100%	75%	90%	95%
Total	100%	87%	94%	96%

Table 1: Results of Automatic Scoring Prototype

ERROR ANALYSIS

An error analysis of the data indicated the following two error categories that reflected a methodological problem: a) Lexicon Deficiency and b) Concept Grammar Rule Deficiency. These error categories are discussed briefly below. Both error types could be resolved in future research.

Scoring errors can be linked to data entry errors, morphological stripping errors, parser errors, and erroneous rules generated due to misinterpretations of the scoring guide. These errors, however, are peripheral to the underlying methods applied in this study.

Lexical Deficiency

Recall that the lexicon in this study was built from relevant vocabulary in the set of 100 training essays. Therefore, vocabulary which occurs in the test data, but not in the training data was ignored during the process of concept-extraction. This yielded incomplete CSRs, and degraded scoring resulted. For instance, while the core concept of the commonly occurring phrase *one band* is more often than not expressed as *one band*, or *one fragment*, other equivalent expressions existed in the test data some of which did not occur in the training data. From our 185 essays we extracted possible substitutions of the term *one fragment*. These are: *one spot, one band, one inclusive line, one probe, one group, one bond, one segment, one length of nucleotides, one marking, one strand, one solid clump, in one piece, one bar, one mass, one stripe, one bar,* and *one blot.* An even larger sample of essays could contain more alternate word or phrase substitutions than those are listed here. Perhaps, increased coverage for the test data can be achieved if additional standard dictionary sources are used to create a lexicon, in conjunction with the example based method used in this study (Richardson et al., 1993). Corpus-based techniques using domain-specific texts (e.g., Biology textbooks) might also be helpful (Church and Hanks, 1990).

Concept Grammar Rule Deficiency

In our error analysis, we found cases in which information in a test essay was expressed in a novel way that is not represented in the set of concept grammar rules. In these cases, essay scores were degraded. For example, the sentence, *"The action of this mutation would nullify the effect of the site, so the enzyme Y would not affect the site of the mutation."* is expressed uniquely, as compared to its paraphrases in the training set. This response says in a somewhat roundabout way that *due to the mutation, the enzyme will not recognize the site and will not cut the DNA at this point.* No rule was found to match the CSR generated for this test response.

SUMMARY AND CONCLUSIONS

This prototype scoring system for AP Biology essays successfully scored the *Excellent* and *Poor* essays with 87% exact agreement with human grader scores. For the same set of essays, there was 94% agreement between the computer scores and human rater scores for exact or adjacent scores. The preprocessing steps required for automated scoring are mostly automated. Manual processes, such as lexicon development could be automated in the future using standard context-based, word distribution methods (Smadja, 1993), or other corpus-based techniques. The error analysis from this study suggests that dictionary-based methods, combined with our current example-based approach, might effectively help to expand the lexicon). Such methods could broaden the lexicon and reduce the dependencies on training data vocabulary. The automation of the fine-tuned CSRs will require more research. A fully automated process would be optimal with regard to time and cost savings. Work at the discourse level will have to be done to deal with more sophisticated responses which are currently treated as falling outside of the norm.

Perhaps the most attractive feature of this system in a testing environment is that it is defensible. The representation used in the system denotes the content of essay responses based on lexical meanings and their relationship to syntactic structure. The computer-based scores reflect the computer-based analysis of the response content, and how it compares to the scoring guide developed by human experts. Information generated by the system which denotes response

content can be used to generate useful diagnostic feedback to examinees.

Since our methods explicitly analyze the content of text, these or similar methods could be applied in a variety of testing, training or information retrieval tasks. For instance, these natural language processing techniques could be used for World Wide Web-based queries, especially with regard to scientific subject matter or other material producing constrained natural language text.

ACKNOWLEDGMENTS

We are grateful to the College Board for support of this project. We are thankful to Altamese Jackenthal for her contributions to this project. We are also grateful to Mary Dee Harris and two anonymous reviewers for helpful comments and suggestions on earlier versions of this paper.

References

Burstein, Jill C., Randy M. Kaplan, Susanne Wolff and Chi Lu. (1996). Using Lexical Semantic Techniques to Classify Free Responses. *Proceedings from the SIGLEX96 Workshop*, ACL, University of California, Santa Cruz.

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.

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

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..

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

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

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.

Wilks, Y., D. Fass, C. Guo, J. McDonald, T. Plate, and B. Slator. (1992). Providing Machine Tractable Dictionary Tools. In J. Pustejovsky (Ed), *Semantics and the Lexicon*, Kluwer Academic Publishers, Boston, MA.

Appendix 1: Sample Rule Matches for a Scored Essay

Part A:
*"The cleaved **DNA** is then placed in a gel electrophoresis box that has a positive and a **negative** end to it."*
Rubric category : CHARGE
Rubric Rule:XP:[DNA],XP:[NEGATIVE]

*"The **longer**, heavier bands would **move** the **least** and the smaller lighter bands would move the most and farther from the starting point."*
Rubric category: RATE/SIZE
Rubric Rule:XP: [LARGE_SIZE],XP:[MOVE,LESS]

Part B:
*"If the DNA was digested with only enzyme X then there would be **4** separate bands that would develop."*
Rubric category :Treatment I
Rubric Rule:XP:[FOUR]

*" If the DNA was digested only with enzyme Y then **two fragments** or RFLP's would be visible."*
Rubric Category: Treatment II
Rubric Rule:XP:[TWO,FRAGMENT]

*"If the DNA was digested with both the X and the Y enzyme then there would be **5** RFLP's of 400 **base pairs**, 500 base pairs, 1,200 base pairs, 1,300 b.p and 1,500 b.p."*
Rubric category : Treatment III
Rubric Rule: XP:[FIVE,FRAGMENT]

*"If the DNA was undigested then we would find no RFLP's and, as a result, there would be **no banding** that would occur."*
Rubric category: Treatment IV
Rubric Rule:XP:[NOT,FRAGMENT]

Parts C1 and C2
*"Restriction enzymes are types of proteins which recognize certain recognition sites along the DNA sequence and **cleave** the **DNA** at that end."*
Rubric category RECOGNITION
Rule:XP:[CUT,DNA]

*"Therefore, there would be **no cut** at that location and no RFLP produced at the Y recognition **site**."*
Rubric Category
Rule:XP:[NOT],XP:[CUT],XP:[SITE]