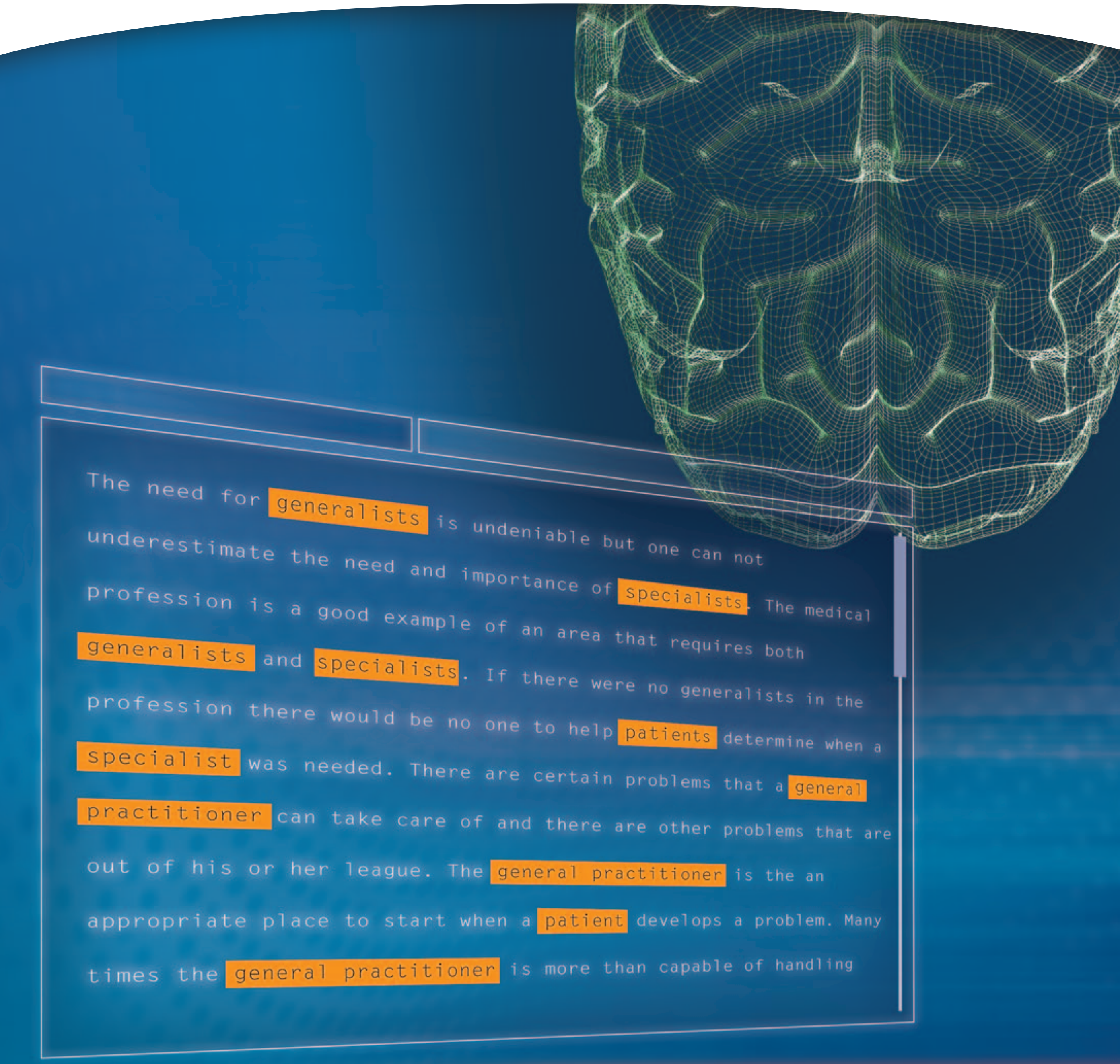
Listening. Learning. Leading.®



The need for generalists is undeniable but one can not underestimate the need and importance of specialists. The medical profession is a good example of an area that requires both generalists and specialists. If there were no generalists in the profession there would be no one to help patients determine when a specialist was needed. There are certain problems that a general practitioner can take care of and there are other problems that are out of his or her league. The general practitioner is the an appropriate place to start when a patient develops a problem. Many times the general practitioner is more than capable of handling

ETS Research Spotlight

# Automated Scoring: Using Entity-Based Features to Model Coherence in Student Essays

ISSUE 5, AUGUST 2011

# Foreword

ETS Research Spotlight

ISSUE 5

AUGUST 2011

**Editor:** Jeff Johnson

**Contributors:** Jill Burstein, Joel Tetreault, Slava Andreyev

**Reviewers:** Dan Eignor, Derrick Higgins, Paul Deane, Ed Shea

**Copy Editor:** Eileen Kerrigan

**Layout:** Sally Acquaviva

**Cover Design:** George Barrett

At ETS, our researchers have extensive experience in natural language processing (NLP) — a field that applies principles of computational linguistics and computer science to the task of creating computer applications that analyze human language.

NLP is the basis for applications that we are developing to address the increasing demand for evaluating open-ended or *constructed-response* tasks, not only on tests, but also in instructional settings. Examples of constructed-response tasks include short written answers, essays, recorded speech, and some math problems.

This issue's Featured Research Synopsis summarizes an example of research we are conducting in this area: a recent paper by Principal Research Scientist Jill Burstein, Research Scientist Joel Tetreault, and Senior Software Developer Slava Andreyev. In their work, Burstein, Tetreault, and Andreyev examined an approach to improving the set of characteristics that automated scoring systems are able to analyze when evaluating student writing.

Such research has benefits not only for the testing programs that use the *e-rater*® Scoring Engine — ETS's automated system for scoring essay-length writing; it also has instructional benefits, potentially leading to tools that help teachers maximize their use of classroom time and resources. The *Criterion*® Online Writing Evaluation Service is an example of one such system, and our researchers are constantly working on other capabilities.

Pages 5–8 of this issue contain a brief list of recent publications that our staff members have authored related to the subject of NLP, as well as a list of selected ongoing research projects that advance the capability of NLP applications. For example, we are working to:

- evaluate the quality and sophistication of speech;
- identify ideas and evaluate the logical relationships among ideas in text;
- base evaluations of writing quality on a more complete construct definition; and
- help teachers prepare instructional materials to support English language learners.

If you'd like to learn more about the research we conduct in these and other areas to support and improve assessments, visit us on the web at **www.ets.org/research**.

**Ida Lawrence**
Senior Vice President
Research & Development

The views expressed in this report are those of the authors and do not necessarily reflect the views of the officers and trustees of Educational Testing Service.

Copies can be downloaded from:
**www.ets.org/research**
**August 2011**
**Research & Development Division**
**Educational Testing Service**

To comment on any of the articles in this issue, write to ETS Research Communications via email at:
**RDWeb@ets.org**

FEATURED RESEARCH SYNOPSIS

# Using Entity-Based Features to Model Coherence in Student Essays

***Editor's note:** This is a synopsis of a more technical publication that appeared in another forum. The full reference list and technical details regarding this research appeared in the original work, which was:*

1) *Burstein, J., Tetreault, J., & Andreyev, S. (2010). Using entity-based features to model coherence in student essays. In Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics (pp. 681–684). Retrieved from:* ***http://aclweb.org/anthology-new/N/N10/N10-1099.pdf***

## Introduction

Natural language processing, or NLP, is a field that applies principles of computational linguistics and computer science to the task of designing computer applications that interact with human-generated language. NLP technology drives applications such as the *e-rater* essay scoring engine, which ETS uses as a quality control aide in conjunction with the scores that trained human raters assign for essays on the *TOEFL*® test and the *GRE*® General Test. The *e-rater* engine is also used to score essay responses on the American Institute of Certified Public Accountants exam. It is the basis for the essay evaluation component in ETS's *Criterion* Online Writing Evaluation Service, which is available on the web as a classroom instructional tool.

Identifying discourse coherence, or the flow of ideas, in student essays is a target of ongoing research in the field of NLP. The authors of the original paper (ETS scientists Jill Burstein and Joel Tetreault, together with ETS software developer Slava Andreyev) evaluated one approach in particular proposed by Regina Barzilay of the Massachusetts Institute of Technology and Mirella Lapata of the University of Edinburgh.[1] The aim of the research is to identify an algorithm that leads to more reliable ratings of coherence in computer-evaluated essays.

The value of conducting such research is that it can potentially allow NLP developers to create automated scoring applications whose scoring models more closely represent the features that writing scholars might consider when evaluating writing. Assessment researchers would say that successfully incorporating discourse coherence into the current set of features used for automated essay scoring may result in the scoring system having better *construct representation*. In addition, the feature could be integrated into the *Criterion* online writing instruction service to give feedback to students on whether or not their essays are fully coherent. The long-range goal of this avenue of work is to pinpoint the exact places in an essay that may cause confusion to a reader.

[1] Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics, 34*(1), 1-34. Retrieved from: **http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.34.1.1**.

## Method

### *Corpus and Annotation*

The researchers collected approximately 800 essays from three sources: (a) The TOEFL test's Independent essay; (b) The GRE General Test's Issue and Argument essays; and (c) submissions to the *Criterion* service. These *datasets* were the basis for the *corpus*, or the set of linguistic data that an NLP-based application uses as an input when creating the model by which it evaluates human language.

The process of *annotation* also leads to a critical input for NLP applications. In this case, the ETS researchers asked two essay evaluation experts — known as annotators in a research context — to rate each essay on a 2-point scale; they assigned each essay a rating of "H" for "high coherence" or "L" for "low coherence."

### *Implementing Barzilay and Lapata's Algorithm*

In their work, Barzilay and Lapata used an entity-based algorithm to create a model for the types of features present in essays that human raters classified based on different levels of coherence. In computational linguistics, an entity is a noun or a pronoun. The algorithm created a grid classifying all of the entities in an essay according to their role (i.e., *subject*, *object*, or *other*). As part of this algorithm, statistics were computed to quantify the writers' use of entities with respect to such features as their use in adjacent sentences, their presence or absence in different roles, and their *salience*, or frequency, throughout the work.

The ETS researchers used the algorithm to build *coherence models* — the criteria by which the NLP application would determine the degree to which essays contained the same coherence features present in samples that writing experts classified as "high coherence" or "low coherence" essays. The ETS researchers also enhanced the power of the algorithm by using it together with features from the *e-rater* engine that analyze errors in grammar, usage, and mechanics, as well as the writer's style. This allowed the NLP application to not simply "count" the number of entity transitions that might indicate coherence, but also to examine whether the words were used correctly.

## Results and Implications

The ETS researchers found that their experimental NLP application performed best when they combined Barzilay and Lapata's algorithm with NLP features designed to examine writing quality. In this combined configuration, the experimental system was able to more reliably evaluate essay coherence than it was when it analyzed only the coherence features present in the algorithm or only writing quality features such as grammar, usage, mechanics, and style.

> *The ETS researchers found that their experimental system was able to more reliably evaluate essay coherence than it was when it analyzed only the coherence features present in the Barzilay and Lapata algorithm or only writing quality features such as grammar, usage, mechanics, and style.*

The measure for performance was the degree to which the application's labeling of a high-coherence or low-coherence essay agreed with a human expert's application of the same labels to the same essays. For two sets of data — from the TOEFL test and the GRE test — the agreement between the NLP application's coherence rating and the human coherence rating was comparable to the agreement between human raters.

The authors conclude that an entity-based method holds promise for producing reliable discourse coherence ratings for essays. The original paper contains tables detailing the authors' findings.

## References

See the original paper, cited in the Editor's note on page 2, for a list of references that the authors used when conducting and reporting on the research described in this synopsis.

## About the Authors

**Jill Burstein** is a principal research scientist at ETS.

**Joel Tetreault** is a research scientist at ETS.

**Slava Andreyev** is a senior software developer at ETS.

Burstein and Tetreault work in the Applied Research and Development area of the Research & Development division. Andreyev is from the Information Technology area of ETS's Production and Delivery division.

# You May Also Be Interested In . . .

*Below is a select list of recent ETS-authored publications related to automated scoring and natural language processing, as well as some ongoing projects that our staff members are leading or contributing to.*

## PUBLICATIONS

### *Automated Scoring of Speech*

Chen, L., Tetreault, J., & Xi, X. (2010). Towards using structural events to assess non-native speech. *In NAACL-HLT 2010: Proceedings of the 5th Workshop on Building Educational Applications (BEA-5)*. Stroudsburg, PA: Association for Computational Linguistics.

Evanini, K., Higgins, D., & Zechner, K. (2010). Using Amazon Mechanical Turk for transcription of non-native speech. In *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, HLT-NAACL 2010*. Stroudsburg, PA: Association for Computational Linguistics.

Higgins, D., Zechner, K., Xi, X., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language, 25(2),* 282–306.

### *Automated Scoring of Writing Quality*

Burstein, J. & Chodorow, M. (2010). Progress and new directions in technology for automated essay evaluation. In R. Kaplan (ed.), *The Oxford Handbook of Applied Linguistics* (2nd Ed., pp. 487–497). New York: Oxford University Press.

Burstein, J., Tetreault, J., & Andreyev, S. (2010). Using entity-based features to model coherence in student essays. In Human language technologies: *The 2010 Annual Conference of the North American Chapter of the ACL*. Stroudsburg, PA: Association for Computational Linguistics.

Louis, A. & Higgins, D. (2010). Unsupervised prompt expansion for off-topic essay detection. In *Proceedings of the Workshop on Building Educational Applications, HLT-NAACL 2010*. Stroudsburg, PA: Association for Computational Linguistics.

Tetreault, J., Foster, J., & Chodorow, M. (2010). Using parse features for preposition selection and error detection. In *Proceedings of the 2010 Association for Computational Linguistics* (ACL 2010). Stroudsburg, PA: Association for Computational Linguistics.

Tetreault, J., Filatova, E., & Chodorow, M. (2010). Rethinking grammatical error annotation and evaluation with the Amazon Mechanical Turk. In NAACL-HLT: 2010 *Proceedings of the 5th Workshop on Building Educational Applications (BEA-5)*. Stroudsburg, PA: Association for Computational Linguistics.

### *Automated Scoring of Written Content*

Sukkarieh, J. Z., & Bolge, E. (2010). Building a textual entailment suite for evaluating content scoring technologies. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (pp. 3149–3169). Paris: European Language Resources Association.

Sukkarieh, J. Z. (2010). Maximum entropy for the automatic content scoring of free-text responses. In *Proceedings of the 30th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2010)*. Retrieved from: **http://web2.lss.supelec.fr/MaxEnt2010/paper/048.pdf**

### *Automated Scoring of Math Responses*

Bennett, R. E. (2011). *Automated scoring of constructed-response literacy and mathematics items*. Retrieved from: **http://www.acarseries.org/papers/Randy_Bennett-Automated_Scoring.pdf**

### *Educational Applications of Natural Language Processing*

Chodorow, M., Gamon, M., & Tetreault, J. (2011). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing*, 27(3), 419–436.

Han, N.-R., Tetreault, J., Lee, S.-H., & Ha, J.-Y. (2010). Using an error-annotated learned corpus to develop an ESL/EFL error correction system. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (pp. 763–770). Paris: European Language Resources Association.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. In G. Hirst (ed.), *Synthesis lectures on human language technologies* (Vol. 3, pp. 1–134). San Rafael, CA: Morgan & Claypool.

## INITIATIVES AND PROJECTS

### Measuring the Development of Vocabulary and Word Learning to Support Content Area Reading and Learning

*Funding Source: U.S. Department of Education*

The goal of this project is to develop improved methods for measuring vocabulary and word learning in specific subject areas such as social studies or science. It exploits natural language processing technologies to develop detailed maps of the vocabulary demands of different kinds of texts, and uses these maps to define a sampling strategy for assessing breadth of vocabulary knowledge.

### A Technology-Rich Teacher Professional Development Intervention That Supports Content-Based Curriculum Development for English Language Learners

*Funding Source: U.S. Department of Education*

The goal of this project is to develop a technology-rich teacher professional development package to support in-service, content-area teachers in the preparation of accessible content materials to facilitate content comprehension, language skills, and reading comprehension of English language learners. This teacher professional development package includes two components: (1) teacher professional development and (2) instructional authoring. The professional development component will provide teachers with a rich set of teaching strategies targeting the diverse learning needs of English language learners. The instructional authoring component will allow teachers to apply linguistic insight gained in the professional development component to pedagogically adapt content-area curriculum for English language learners.

### Projects Within the ETS Strategic Research Initiatives

*Funding Source: ETS Research Initiatives*

In the context of its Constructed-Response (CR) Scoring Research Initiative, ETS is undertaking foundational research on automated scoring and natural language processing. The purpose of this initiative is to:

- advance the understanding of fundamental characteristics of how expert human raters perceive;
- interpret and evaluate responses to CR items;
- use this understanding to improve human scoring; and
- conceptualize, develop, deploy, and improve automated systems that might complement or in some cases replace the work of human evaluators.

A major goal of the automated scoring portion of this effort includes improving the ability of ETS's automated scoring engines to a) handle meaning in a text in a deeper and more flexible way; and b) more fully address the construct of speaking proficiency and produce more reliable and valid scores for existing high-stakes speaking assessments such as the *TOEFL* test and the *TOEIC* test.

With respect to the improved ability to use automated systems to score text, work in 2011 is focusing on:

- Improvements in the features able to be scored by the *e-rater* and *c-rater*SM scoring engines
- Investigation of ways to detect organizational structure, factual representation, and opinion in a text

With respect to the improved ability to use automated systems to score spoken responses, work in 2011 is focusing on:

- Identification of new features of speech able to be scored by the *SpeechRater*SM service
- Investigation of the performance of a new speech-recognition system

## About ETS

764283

17280

*Listening. Learning. Leading.*®

*www.ets.org*