# International Automated Essay Scoring Systems: An overview

Li Guo
*International Education Center, Beijing Language and Culture University*
*15 Xueyuan Road, Haidian District, Beijing, 100083, P.R. China*

***Abstract* — *The paper introduces the development of automated essay scoring systems, analyses their theoretical foundation, technical path and scoring models, and finally compares the similarities and differences of these systems.***

**Keywords** — *automated essay scoring, PEG, IEA, E-Rater, IntelliMetric$^{TM}$*

## I. BRIEF DEVELOPMENT OF THE AUTOMATED ESSAY SCORING SYSTEMS

Automated essay scoring system is a system which can analyze essays through computer programs, and evaluate and score them according to a statistical model which is already built. The research involves many disciplines, including linguistics and writing research, cognitive psychology, computer science, education measurement, etc.

In the 1960s, Page and his team successfully developed the world's first stable and applicable automated essay scoring system "Project Essay Grade (PEG)".The scoring correlation coefficient between PEG and human raters achieved .70. Actually, before this, the scoring correlation coefficient between two trained raters could only be .60[1]. It is the first time that humans have successfully utilized computers to score essays and realized breakthroughs on automated scoring technology. The further research on automated essay scoring has been intensely restricted due to the limitations of values at that time as well as the slow development of the computer technology. Research in this area thus stopped for quite a long time, and existing achievements remained dormant, without any opportunity to be applied on a large scale.

After mid-1980s, with the progress of computer technology and the emergence of the programming languages, individuals began to have the opportunity to use the computer. Thus the recovery of the automated essay scoring is promoted, which has drawn more attention from researchers. At the same time, the corpus of writing texts was massively employed in the research, such as the text corpus of National Assessment of Educational Progress, 1988.

In the early 1990s, with the development of micro-computer and the internet, individuals can input the essay into the micro-computer via the keyboard directly and submit it through the internet for scoring from various places. Therefore, the automated essay scoring has been revived. Besides that, the diversity in the building theory of the statistical model and the progress of the processing technology of the natural languages provide a strong theoretical and technical support. The researchers has developed a variety of automated essay scoring systems that can be used in the actual rating in a relatively short period of time, such as Intelligent Essay Assessor (IEA), E - rater, IntelliMetric, Bayesian Essay Test Scoring System (BETSY) and so on. They have been widely assisted in the teaching of writing and large-scale exam essay scoring, such as the GMAT, GRE, etc.

## II. THE MAJOR FOREIGN AUTOMATED SCORING SYSTEMS

The foreign automated scoring systems mainly include Project Essay Grading (PEG) systems, Intelligent Essay Assessor (IEA), E-rater, IntelliMetric, etc. These systems are based on different theories, and therefore have their own characteristics.

### *A. Project Essay Grading (PEG)*

PEG is an automated essay scoring system based on the analysis theory of text feature. The analysis theory of text feature is mainly based on the shallow analysis of the linguistic features of the essays, taking the variables easy to quantify as independent variables, such as length, vocabulary and grammar, and the score as dependent variable. The dependent variable is estimated through the statistics of the independent variables, namely, through the establishment of index system of shallow linguistic features, the essays will be graded for the statistical analysis of related linguistic indicators, and finally obtained a score by observing the corresponding relationship with the index system.

Reference [2] argued that, in essay scoring, two categories of variables need to be considered, which he called trins and proxes. Trins refer to the internal variables, such as fluency, wording, grammar, punctuation, etc., which are also a human rater’s concern. Proxes are indicator variables that can be identified in computer scoring, that is, the quantitative embodiment of internal variables. Proxes focus on the correlation of the internal variables, such as establishing the correlation coefficient[3] of fluency (internal variables) and vocabulary (indicator variables). Internal variables (trins) of the essay could be demonstrated by the indicator variables (proxes). For example, the author's writing ability can be inferred from the length of the essay. The complexity of the sentence structure can be measured by the amount of the prepositions, relative pronouns and words with other parts of speech. The author's language ability can be displayed from the changes of the word length [4]. Therefore, essay scoring is to find out trins, compare them with proxes, and then unravel the similarity or approximation. Proxes are feature index established by the shallow analyses of the linguistic characteristics. This shows that the basic principle of the PEG is to establish a statistical regression model with shallow text characteristics (such as length, spelling, etc.) as the independent variables and the proposed essay as the dependent variable. It is quite essential to identify the categories of the independent variables and the weight of each independent variable in the process of establishing the statistical regression model.

In terms of scoring methods, PEG mainly includes two phases: training and scoring. During the training phase, 100 – 400 essays graded by human raters are analyzed to determine the values of related variables, which indicate the values of all the features chosen for each article that has a definite score. Then multiple regression analysis of these values is conducted to calculate regression coefficient (the weight of each features in the score of the essay). A quantitative weight system of feature index is thus established to reflect the relationship between the values of each feature and the final score. During grading phase, the system will calculate relevant variables for each essay input in the computer, input regression prediction formula established based on weight system of feature index, and calculate the final score of the essay.

Reference[5]'s report shows that the value of R for PEG's regression equation is as high as 0.87, highly consistent with human rating. Early version of PEG can only offer an overall score, while later version can gradually analyze other features of the essay, like structure, and style, etc., and also give feedback to the writer[6]. With the development of the Internet technology, PEG researchers also start to focus on online tests, which shows that PEG is an effective assay scoring system.

### B. Intelligent Essay Assessor (IEA)

The success of the IEA system lies in the application of Latent Semantic Analysis (LSA) in directly measuring the quality of an essay. Originally used for information retrieval, LSA is now widely applied in trans-language information retrieval, information filtering, and essay scoring. With statistical computations applied to a large corpus of text, it enables computer to master mathematical expression of the semantic relationship between words and essay. Its basic idea is that: the idea of an essay lies in the words used, that is to say, all the words in an essay contribute to the general idea, which wouldn't be affected even when one word changes. To put it another way, two essays containing different words may convey the same idea. The general idea of an essay is the sum of the meanings of its words, from which semantic features can be extracted. Simply, meaning of word No. 1 + meaning of word No. 2 + … + meaning of word No. n = meaning of the essay [7]. And its basic operation is to project the essays in a high-dimensional vector space model to a low-dimensional latent semantic space, achieved by singular value decomposition (SVD) of term-document matrix[8].

In the specific application of LSA theory, individuals analyze statistically a large number of texts of an existing area of knowledge, establish a multidimensional semantic space model between words and essays in this area, judge semantic relevance and similarities among different essays by comparing their corresponding positions in semantic space model, and then predict semantic similarities between two essays. In the application of LSA to essay scoring, the technical path is to project intended score as words contained to mathematical form that can represent the meaning of essay, compare it with essays already graded by human raters in terms of concept relevance and content, then the score is provided. The whole process can be divided to three steps:

First, extract and use texts of an area of knowledge. Sources of the text data can be graded essays, expert model essays, model essays of this area of knowledge, or a part of ungraded essays.

Second, establish a semantic space model in which words correspond to documents by conducting statistical analysis of text data already extracted with SVD.

Third, compare essays to be graded and the semantic space model stated above in terms of concept relevance and content, then the score can be calculated.

In terms of operation, some researchers believe that compared with the number of training samples needed by other automated essay scoring system, which is at least 300-500, IEA's need on this is rather small, which is only 100 samples.

Developers of IEA argue that it is the only program that can measure semantics and content of essays[9], and its score is highly correlated with that of human raters, with the coefficient reaching 0.85[10].

### C. E-rater

E-rater was originally developed by Educational Testing Service (ETS) in the 1990s for evaluating the quality of GMAT essays. It judges the quality of an essay by analyzing specific vocabulary and syntax with natural language processing (NLP) of artificial intelligence (Burstein, 2003; Kukich, 2000). It mainly comprises five modules, of which three modules that are based on syntax, discourse structure and topical analysis of NLP are used to extract features that reflected writing qualities specified in scoring criteria. In syntax module, syntactic analyzer is used to tag each word for part-of-speech, find phrases based on this, assembles phrases into trees based on subcategorization information for verbs, and then different clauses are identified. In discourse structure module, discourse analyzer annotates each essay according to a discourse classification schema and partition essays into separate arguments. In topical analysis module, analyses are done at the level of the essay and the argument. Training essays are converted into vectors of word frequencies, and the frequencies are then transformed into word weights. Based on these weight vectors, cosine similarity analysis between test and training vectors is conducted. 67 characteristics are extracted from these three modules. The fourth module is used to build models, select variables, establish regression models, choose and weigh features predictive of essay scoring, screen 67 characteristics and establish regression equation. The fifth module is used to extract values of features for essays to be graded, substitute them into the regression equation, and compute the final score.

### *D. IntelliMetric$^{TM}$*

Officially put into commercial use in 1998, IntelliMetric$^{TM}$ was an intelligent scoring system that combines artificial intelligence, NLP and statistical technology, internalizes the pooled wisdom of many expert raters, and can emulate the process carried out by human scorers. Its core technologies are CogniSearch and Quantum Reasoning of Vantage Learning. After analyzing more than 300 features of essays at the level of semantics, syntax and text, it divides them into five categories: focus and unity (features pointing towards cohesiveness and consistency in purpose and main idea); development and elaboration (features of text looking at the breadth of content and support for concepts advanced); organization and structure (features targeted at the logic of discourse including transitional fluidity and relationships among parts of the response); sentence structure (features targeted at sentence complexity and variety); and mechanics and conventions (features examining conformance to conventions of edited American English). A total of more than 300 features are identified to be substituted into the model for scoring. The steps are: First, train the system with essays already graded by human raters; these essays will provide the system with scoring dimensions and expert wisdom. After training, the system will learn about the relationship between scoring dimensions, essay features and score. Second, model equation is determined by conducting reliability analysis test. Third, score the essays with model equation. The consistency with expert raters can reach 97% - 99%.

## III. Comparison of the Four Systems

Automated essay scoring systems can’t evaluate directly the inherent quality of essays; what they do is to predict essay score with correlation coefficient of inherent quality. Although the theoretical basis of PEG, IEA, E-rater and IntelliMetric are quite different from each other, they all need different numbers of human graded essays as training set to train the system, establish their own feature systems, compare them with essays to be graded, and finally compute the score. Table I is a comparison of the four systems.

All of the four automated essay scoring systems have their own advantages and disadvantages. As the first system of such kind, PEG focuses on the analysis of text features, while ignoring content. On the contrary, the second system IEA emphasizes that scoring of content is important. Starting from the third group of systems, E-rater and IntelliMetric, researchers begin to pay attention to both text feature analysis and scoring of the content. Meanwhile, their research methods and vision for test are not completely the same. Table II analyses their features respectively.

Table I Comparison of Four Systems

| Name | Inventor and Time of Invention | Theoretical Basis | Focus of Scoring | Number of Training Set | Relevance / Consistency |
|---|---|---|---|---|---|
| PEG | Page，1966 | Text features of form | Text features | 100-400 | R=0.87 |
| IEA | Landauer et al.1997 | Latent Semantic Analysis | Semantic features | 100-300 | R=0.85 |

| | | | | | |
|---|---|---|---|---|---|
| E-rater | Burstein et al. 1998 | Natural Language Processing | Text and semantic features | 465 | >97% |
| IntelliMetric$^{TM}$ | Elliot et al.1998 | Artificial Intelligence | Text and semantic features | 300 | 97%-99% |

Table II Advantages and Disadvantages of the Four Systems

| **Name** | **Methods and Paths** | **Advantages** | **Disadvantages** |
|---|---|---|---|
| PEG | Multiple regression analysis;<br>Train the system and test reliability by human scoring | Text analysis | Only shallow text features are analyzed, content and discourse structure are ignored. |
| IEA | Train the system and test reliability by human scoring | Semantic analysis | There is a lack of analysis of text features; discourse structure is ignored. |
| E-rater | Multiple regression analysis;<br>Train the system and test reliability by human scoring;<br>Establish different modules | Establish different modules; take into consideration the analysis of discourse structure | Analysis for content and language quality is not enough; discourse analysis is limited to superficial features. |
| IntelliMetric$^{TM}$ | Train the system and test reliability by human scoring; establish different modules | More than 300 features at the level of semantics, syntax and discourse | The structure is too complicated. |

## IV. CONCLUSION

After decades' development, the international mainstream automated scoring systems have formed a relatively favorable system structure model and algorithm. Whereas, the latest research achievements are hardly exposed to the public because of the protection of intellectual property rights. In other words, the current research achievements shown in front of the public may not be on behalf of the latest research level. However, thought of the research, as well as the technical path, still can be traced from the public achievements, and eventually developed their own automated essay scoring system. Now quite a few commercial English automated essay scoring systems are developed in the domestic market, like www.pigai.org and writing.bingoenglish.com. With the thorough research of IT and artificial intelligence, automated essay scoring will be recognized by more people and be promoted on a large scale to further assist students in improving their writing.

## REFERENCES

[1] Page, E.B. Project Essay Grading: PEG. In M.D. Shermis& J. Burstein (Eds), Automated essay grading: A cross-disciplinary perspective(pp.43-54). Mahwah, NJ: Lawrence Erlbaum Associates, Inc, 2003.

[2] Page, E. B. The use of the computer in analyzing student essays. *International Review of Education*, 14, 210–225, 1968.

[3] Dikli, S. An Overview of Automated Scoring of Essays. *Journal of Technology Learning & Assessment*, 5(1),4-34, 2006.

[4] Ge Shili & Chen Xiaoxiao. An Overview of Current Automated Essay Scoring Technique. *CAFLE* (05), 25-29, 2007.

[5] Page, E. B. Computer grading of student prose, using modem concepts and software. *Journal of Experimental Education*, 62(2), 127-142, 1994.

[6] Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S, "Trait ratings forautomated essay grading", in *Measurementin Education*, Montreal, Canada, 1999.

[7] Landauer, T. K., Laham, D., & Foltz, P. W. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M.D. Shermis& J. Burstein (Eds), *Automated essay grading: A cross-disciplinary perspective* (pp.87-112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc, 2003.

[8] Cai Li, Peng Xingyuan & Zhao Jun. Research on Assisted Scoring System for Chinese Proficiency Test for Minorities. *Journal of Chinese Information Processing* (05), 120-126, 2011.

[9] Liang Maocheng & Wen Qiufang. A Critical Review and Implications of Some Automated Essay Scoring Systems. *CAFLE* (05), 18-24, 2007.

[10] Landauer, T. K., &Psotka, J. Simulating Text Understanding for Educational Applications with Latent Semantic Analysis: Introduction to LSA. *Interactive Learning Environments*, volume 8(2), 73-86, 2000.