# DESIGN OF AN AUTOMATED ESSAY GRADING (AEG) SYSTEM IN INDIAN CONTEXT

By

SIDDHARTHA GHOSH *

SAMEEN S FATIMA**

*ABSTRACT*

*Automated essay grading or scoring systems are no more a myth, but they are a reality. As on today, the human written (not hand written) essays are corrected not only by examiners / teachers but also by machines. The TOEFL exam is one of the best examples of this application. The students' essays are evaluated both by human and web based automated essay grading system and then the average is taken. Many researchers consider essays as the most useful tool to assess learning outcomes, implying the ability to recall, organize and integrate ideas, the ability to supply, merely than identify interpretation and application of data. Automated Writing Evaluation Systems, also known as Automated Essay Assessors, might provide precisely the platform we need to explicate many of the features those characterize good and bad writing and many of the linguistic, cognitive and other skills those underline the human capability for both reading and writing. They can also provide time-to-time feedback to the writers/students which can improve their writing skill. A meticulous research of last couple of years has helped us to understand the existing systems which are based on AI & Machine Learning techniques and finding the loopholes and at the end to propose a system, which will work under Indian context, presently for English language influenced by local languages. Currently most of the essay grading systems are used for grading pure English essays or essays written in pure European languages. In India we have almost 21 recognized languages and influence of these local languages in English, is very much here. Newspapers in sometimes print like "Now the time has come to say 'albida' (good bye) to monsoon". Due to the influence of local languages and English written by non-native English speakers (ie. Indians) the result of TOEFL exams has shown lower scores against Indian students (also Asian students). This paper focuses on the existing automated essay grading systems, basic technologies behind them and proposes a new framework to overcome the problems of influence of local Indian languages in English essays while correcting and by providing proper feedback to the writers.*

## INTRODUCTION

Evaluation and Grading play a central role in the educational process. The interest in the development and in use of *Computer-based Assessment Systems* (CbAS) has grown exponentially in the last few years, due to the increase in the number of students attending universities, and the possibilities provided by e-learning approaches to asynchronous and ubiquitous education. Presently more than forty commercial CbAS are currently available on the market. Most of those tools are based on the use of the so-called objective-type questions: i.e. multiple choice, multiple answer, short answer, selection/association, hot spot and visual identification. Most researchers in this field agree on the notion that some aspects of complex achievement are difficult to measure using objective-type questions. Learning outcomes implying the ability to recall, organize and integrate ideas, the ability to express oneself in writing and the ability to supply, merely than identify interpretation and application of data, requires less structuring of response than that imposed by objective test items (Gronlund, 1985). It is in the measurement of such outcomes, corresponding to the higher levels of the Bloom's (1956) taxonomy (namely evaluation and synthesis) that the essay question serves its most useful purpose. One of the difficulties of grading essays is the subjectivity, or atleast the perceived subjectivity, of the grading process. Many researchers claim that the subjective nature of essay assessment leads to variation in grades awarded by different human assessors, which is perceived by students as a great source of unfairness.

Furthermore essay grading is a time consuming activity. It is found that about 30% of teachers' time is devoted to marking. A system for automated assessment would at least be consistent in the way it scores essays, and enormous cost and time savings could be achieved if the system can be shown to grade essays within the range of those awarded by human assessors. Furthermore using computers to increase our understanding of the textual features and cognitive skills involved in the creation and in the comprehension of written texts, provide a number of benefits to the educational community.

Purpose of this article  is to present a new concept over the existing ones, through which we can overcome the problem of influence of local Indian languages in English essays. The system can do the grading of English essays as well as it can also provide sufficient feedback so that the students/user can understand  the basic errors (spelling, grammar, sentence formation etc.) made by them and whether their essay is influenced by local language or not, and how to overcome all these problems. The paper also discusses the current approaches to the automated assessment of essays (English Essays) and utilizes this as a foundation for the new framework. Thus, in the next section, functioning of some of the following important automated grading systems will be discussed: Project Essay Grade (PEG), Intelligent Essay Assessor (IEA), Educational Testing service I, Electronic Essay Rater (ERater), C-Rater, BETSY, Intelligent Essay Marking System, SEAR, Paperless School, free text Marking Engine and Automark. All these systems are currently available either as commercial systems or as the result of research in this field. In the later sections, the concept of the new system is described.

## 1. Various automated essay-grading systems

Automated scoring capabilities are especially important in the realm of essay writing. Essay tests are a classic example of a constructed-response task where students are given a particular topic (also called a prompt) to write about. The essays are generally evaluated for their writing quality. Surprisingly for many, automated essay scoring (AES) has been a real and viable alternative, and complement to human scoring for many years. As early as 1966, Page showed that an automated "rater" is indistinguishable from human raters (Page, 1966). In the 1990's more systems were developed; the most prominent systems are the Intelligent Essay Assessor (Landauer, Foltz, & Laham, 1998), Intellimetric (Elliot, 2001), a new version of the Project Essay Grade (PEG, Page, 1994), and e-rater (Burstein et al., 1998).

Ellis Page set the stage for automated writing evaluation (Figure 1).  Recognizing the heavy demand placed on teachers and large-scale testing programs in evaluating student essays, Page developed an automated essay-grading system called Project Essay Grader (PEG). He started with a set of student essays that teachers had already graded. He then experimented with a variety of automatically extractable textual features and applied multiple linear regressions to determine an optimal combination of weighted features that best predicted the teachers' grades. His system could then score other essays using the same set of weighted features. In the 1960s, the kinds of features someone could automatically extract from text were limited to surface

| Pioneering Writing evaluation research | | Recent Essay grading research | | Operational systems | Current ETS research | Future research & application | |
|---|---|---|---|---|---|---|---|
| PEG page | Writer's Workbench MacDonald et al. | PEG page | Computer Analysis of Essay content Burstein et al. Intelligent Essay Assessor Landauer et al. PED Page & Petersen | PEG page e-rater ETS Latent semantic analysis knowledge analysis technologies criterion ETS Technologies | Writing diagnosis chodorow &Leacock Mitsakaki& Kulich Burstien & Marcu | Short answer scoring Leacock& chodorrow Hirchman et al. Breck at al. | Questioning answering system Light et al. Verbal test Creation tools students-centered instruction of systems Erater-v.2 |
| 1966 -1968 | 1982 | 1994-1995 | 1997 | 1998-2000 | 2000 | 2000-2006 | |

Figure 1. A timeline of research developments in writing evaluation

features. Some of the most predictive features Page found included average word length, essay length in words, number of commas, number of prepositions, and number of uncommon words, the latter being negatively correlated with essay scores.

In the early 1980s, the Writer's Workbench tool (WWB) set took a first step toward this goal. WWB was not an essay-scoring system. Instead, it aimed to provide helpful feedback to writers about spelling, diction, and readability. In addition to its spelling program one of the first spelling checkers  WWB, included a diction program that automatically flagged commonly misused and pretentious words, such as ‘irregardless’ and ‘utilize’. It also included programs for computing some standard readability measures based on word, syllable, and sentence counts, so in the process it flagged lengthy sentences as potentially problematic. Although WWB programs barely scratched the surface of text, they were a step in the right direction for the automated analysis of writing quality.

In February 1999, E-rater became fully operational within ETS's Online Scoring Network for scoring GMAT essays. For low-stakes writing-evaluation applications, such as a Web-based practice essay system, a single reading by an automated system is often acceptable and economically preferable. The new version of e-rater (V.2) is different from other automated essay scoring systems in several important respects. The main innovation of e-rater V.2 is a small, intuitive, and meaningful set of features used for scoring; a single scoring model and standards can be used across all prompts of an assessment; modeling procedures that are transparent and flexible, and can be based entirely on expert judgment.

Figure 2. shows a popular, common frame work of the automated essay grading systems. Most of the modern systems train the system with almost thousands of pre-assessed essays (corpus). Then, once the essay input is given, it gives the grade as well as a proper feedback to improve.  Hence some of these systems can be used for self-learning by students as well as by the teachers or institutes for grading huge amount of essays. But recently , from the year of 2007 the internationally recognized TOEFL exam gives the grade to the students' essays as a combination of human and machine assessment.

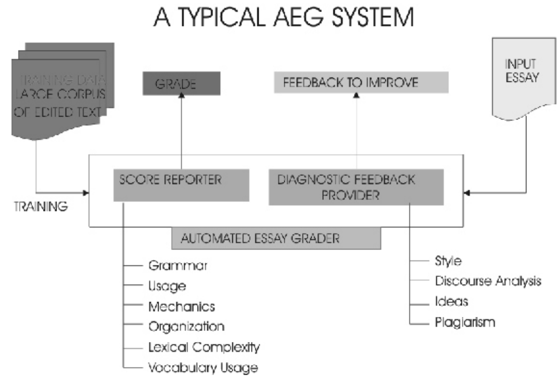

Figure 2. A common framework for the existing Automated Essay Grading Systems

### 2. How the AEG systems work?

AEG systems are a combination of any two, three or all the techniques mentioned here: NLP (Natural Language Processing), Statistics, Artificial Intelligence (Machine Learning), Linguistics and Web Technologies, Text Categorization, annotated large corpora etc. It must be noted that seven out of ten most popular systems are based on the use of Natural Language Processing tools, which in some cases are complemented with statistical based approaches. How does it comes under Artificial Intelligence?  When a  machine can grade human written essays, which requires some expertise, then it can be called as Artificial Intelligence. Because, the commonly available systems cannot perform that task. Text categorization is the problem of assigning predefined categories to free text document. The idea of automated essay grading, based on text categorization techniques, text complexity features and linear regression methods was first explored by Larkey (1998). The underlying idea of this approach relies on training of binary classifiers to distinguish “good” from “bad” essays and on using the scores produced by the classifiers to rank essays and assign grades to them. Several standard text categorization techniques are used to fulfill this goal: first, independent Bayesian classifiers allow assigning probabilities to document and estimate the likelihood that they belong to specific classes; then, an analysis of the occurrence of certain words in the documents is carried out,  and a k-nearest neighbor technique is used to find those essays closest to a sample of human graded essays; finally, eleven text complexity features are used to assess the style of the essays. Larkey conducted a number of regression trials, using different combinations of components. She also used a number of essay sets, including essays on social studies, where content was the primary interest, and essay on general opinion where style was the main criteria for assessment.

A growing number of statistical learning methods have been applied to solve the problem of automated text categorization in the last few years, including regression models, nearest neighbor classifiers, Bayes belief networks, decision trees, rule learning algorithms, neural networks and inductive learning systems (Ying, 1997). This growing number of available methods is raising the need for cross method evaluation.

But the most relevant problem in the field of automated essay grading is the difficulty of obtaining a large corpus of essays (Christie, 2003; Larkey, 2003), each with its own grade on which experts agree. Such a collection, along with the definition of common performance evaluation criteria, could be used as a test bed for a standardized comparison of different automated grading systems. Moreover, these text sources can be used to apply to automated essay grading to the machine learning algorithms, well known in NLP research field, which consist of two steps: a training phase, in which the grading rules are acquired using various algorithms, and a testing phase, in which the rules gathered in the first step are used to determine the most probable grade for a particular essay. The weakness of these methods is the lack of a widely available collection of documents, because their performances are strongly affected by the size of the collection. A larger set of documents will enable the acquisition of a larger set of rules during the training phase, thus a higher accuracy in grading. A major part of these techniques, giving training to the systems and in later stage making the systems to learn from new essays or experience is nothing but machine learning.

The feature set used with some modern AEG systems include measures of grammar, usage, mechanics, style, organization, development, lexical complexity, and prompt-specific vocabulary usage. This feature set is based in part on the natural language processing foundation that provides the instructional feedback to students who are writing essays. In some cases a web-based service evaluates a student's writing skill and provides instantaneous score reporting and diagnostic feedback. The score engine or score reporter (Figure 2.) provides score reporting. The diagnostic feedback is based on a suite of programs (writing analysis tools) that identify the essay's discourse structure, recognize undesirable stylistic features, and evaluate and provide feedback on errors in grammar, usage, and mechanics. The writing analysis tools identify five main types of grammar, usage, and mechanics errors, agreement errors, verb formation errors, wrong word use, missing punctuation, and typographical errors. The approach to detecting violations of general English grammar is corpus based and statistical, and can be explained as follows: In case of corpus based systems, the system is trained on a large corpus of edited text.

### 3. Problems with the present systems under Indian context

It has been found that most of the popular AEG systems are made to grade English essays and they are easy to follow. Systems developed in non-English languages are not popular and not understandable for everyone. This article shows that while a system grades an English essay, it considers the influence of local languages as Error. Hence the following two sentences will show error once they are evaluated by machine as well as by a native-English speakers. *EX. 1) Prime Minister Manmohan Singh* Garu *has visited Osmania University.* In this sentence 'Garu' is a pure Telugu word and used in English newspapers published form Andhra Pradesh. *2) Hyderabad says* "albida" *to monsoon.* Here 'Albida' is an Urdu word and very much used in English newspapers coming out from Lucknow and Hyderabad. Influence of local languages like Maharashtra, Assami, Bengali, Tamil, etc have greatly influenced English in India and no one considers them as Error. Whereas in the view of native speakers of English, they are wrong. Infact a good number of Hindi loan words got chance to be included in

Oxford dictionary. Research shows, that the present AEG systems illustrate 10 - 15% lower score while using Indian English text as Input. In a broader form it can be mentioned that the English spoken and written by non-native English people (i.e. - Asians) are very much influenced by local languages. India is a multilingual country with as many as 22 scheduled languages and only 5% of the population is able to understand English. Hence the goal of the study is to develop a framework for an AEG system, which can be used for correcting essays written in Indian Languages, and also to teach how to write better English Essays, and a standard framework has been proposed to develop any Automated Essay Grading System under Indian context. This model can be executed or the software can be build as per the requirement, for example it can be designed according to the specific regions of India, where the system is going to be used. Because while writing English the students of Andhra Pradesh are not influenced by Bengali or Tamil, but by Telugu. Hence a single system will not be able to solve the problem. But this framework can be used as a benchmark to develop the other AEG systems under Indian context. The framework follows IEEE Std. 1471 2000, which is about "IEEE Recommended Practice for Architectural Description of Software Intensive Systems".

### 4. Proposed framework

Under the above circumstances, a need of a specialized AEG system was felt very much. Hence a new framework is proposed, where the system will have the capability of identifying the local languages (Indian) presented in the submitted essay and it will also find out the effect of these words. It will also help the students to resubmit the essay with corrections where the students will be asked to re-enter the similar words in English, instead of the local languages. Their essays will be graded as they have entered the equivalent English words by their own. For the instructors or teachers it will also give a proper score-card by mentioning the extent to which the essay is influenced by the local languages and the no of local words present, number of times corrections made by students (i.e. they can be given two or three chances to enter equivalent English words for local words (i.e. albida = good bye). The above-mentioned action is a part of the scoring engine. These functionalities are added as a new functional module in the scoring engine or score reporter.

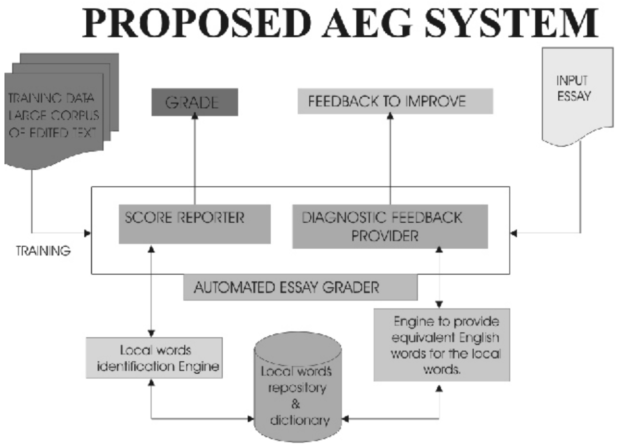
PROPOSED AEG SYSTEM
TRAINING DATA LARGE CORPUS OF EDITED TEXT
GRADE
FEEDBACK TO IMPROVE
INPUT ESSAY
TRAINING
SCORE REPORTER
DIAGNOSTIC FEEDBACK PROVIDER
AUTOMATED ESSAY GRADER
Local words identification Engine
Local words repository & dictionary
Engine to provide equivalent English words for the local words.

Figure 3. Proposed framework of the AEG system with local language engines.

The feed back module is also supported with a 'local language' engine which helps the students to provide proper feed back and development notes along with the instruction to improve English Grammatical mistakes, and notify on the use of too many weak or common words etc. This engine will be very much useful in the learning stage. At the very beginning this engine will identify the local language presented in the written essay. Then it will give a chance to the students to overcome this problem by providing equivalent English words by their own. Then it will show them the projected score with number of general (English) errors and presence of number of local languages and what they are. For the remaining local words used in the essay, the system will then suggest equivalent English words with the synonyms. Now the students get a chance to substitute remaining local words and phrases with suggested English words. After submission they get the final projected score. Hence these engines help the students to learn better English. To make these engines effective, the system is trained by the author with a good number of local words, that are very much used in normal English (spoken English, news paper English). To make a proper collection of local words, the local English news papers are used as a source. As for example, to make the engine working in Andhra Pradesh, it is trained on collection of local words used in the news papers like Deccan Chronicle, Hindu (AP edition), Times of India (AP edition) etc., collected over last couple of years. It is found that this specific region's English is influenced by Telugu and Hyderabadi Hindi (a good mixing of Hindi and Urdu).

## Conclusion

In his paper 'Region Effects in AEG & human discrepancies of TOEFL score' Attali (2005) mentioned that Asian Students show higher organization scores and poor grammar usage and mechanics scores compared to other students. Moreover, local languages have influenced them to a large extent. Serious work in the area of AEG can bring significant changes in this direction and also can give a new shape to Indian Text Categorization & Machine Learning research work.

## Future plans

In near future the following things will be taken into consideration so that some solutions can be provided. They are: Solution for machine translated essays (how to recognize them?), Capturing the mental status of the student writing the essay (psychometric models will be considered), Detection of Anomalous Essays. As an overall research work our focus also will be on the issue- Can we really develop a NLP based Essay Grading System?

## References

[1]. **Bloom, B.S. (1956).** Taxonomy of educational objectives: The classification of educational goals. Handbook I, Cognitive domain. New York, Toronto: Longmans, Green.

[2]. **Burstein, J., Kukich, K., Wolff, S., Chi, L., & Chodorow M. (1998).** Enriching automated essay scoring using discourse marking. Proceedings of the Workshop on Discourse Relations and Discourse Marking, Annual Meeting of the Association of Computational Linguistics, Montreal, Canada.

[3]. **Burstein, J., Leacock, C., & Swartz, R. (2001).** Automated evaluation of essay and short answers. In M. Danson (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University, Loughborough, UK.

[4]. **Christie, J. R. (1999).** Automated essay marking-for both style and content. In M. Danson (Ed.), Proceedings of the Third Annual Computer Assisted Assessment Conference, Loughborough University, Loughborough, UK.

[5]. **Christie, J. R. (2003).** Email communication with author. 14th April. Cucchiarelli, A., Faggioli, E., & Velardi, P. (2000). Will very large corpora play for semantic disambiguation the role that massive computing power is playing for other AI-hard problems? 2nd. Conference on Language Resources and Evaluation (LREC), Athens, Greece.

[6]. **Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman R. A. (1990).** Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

[7]. **de Oliveira, P.C.F., Ahmad, K., & Gillam, L. (2002).** A financial news summarization system based on lexical cohesion. Proceedings of the International Conference on Terminology and Knowledge Engineering, Nancy, France.

[8]. **E.B. Page,** "The Use of the Computer in Analyzing Student Essays," Int'l Rev. Education, Vol. 14, 1968, pp. 210225.

[9]. **E.J. Breck et al.,** "How to Evaluate Your Question Answering System Every Day … and Still Get Real Work Done," Proc. LREC-2000, Linguistic Resources in Education Conf., Athens, Greece, 2000.

[10]. **Grondlund, N. E. (1985).** Measurement and evaluation in teaching. New York: Macmillan.

[11]. **Hearst, M. (2000).** The debate on automated essay grading. IEEE Intelligent Systems, 15(5), 22-37, IEEE CS Press. Honan, W. (1999, January 27). High tech comes to the classroom: Machines that grade essay. New York Times.

[12]. **Jerrams-Smith, J., Soh, V., & Callear D. (2001).** Bridging gaps in computerized assessment of texts. Proceedings of the International Conference on Advanced Learning Technologies, 139-140, IEEE.

[13]. **Laham, D. & Foltz, P. W. (2000).** The intelligent essay assessor. In T.K. Landauer (Ed.), IEEE Intelligent Systems,

2000.

[14]. **Landauer, T. K., Foltz, P. W., & Laham D. (1998).** An introduction to latent semantic analysis. Discourse Processes, 25. Retreived from http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

[15]. **Larkey, L. S. (1998).** Automatic essay grading using text categorization techniques. In Proceedings of the 21st ACM/SIGIR (SIGIR-98), 90-96. ACM.

[16]. **Larkey, L. S. (2003).** Email communication with author. 15th April. Mason, O. & Grove-Stephenson, I. (2002). Automated free text marking with paperless school. In M. Danson (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University, Loughborough, UK.

[17]. **Ming, P.Y., Mikhailov, A.A., & Kuan, T.L. (2000).** Intelligent essay marking system. In C. Cheers (Ed.), Learners Together,Feb. 2000, NgeeANN Polytechnic, Singapore. http://ipdweb.np.edu.sg/lt/feb00/intelligent_essay_marking.pdf

[18]. **Mitchell, T., Russel, T., Broomhead, P., & Aldridge N. (2002).** Towards robust computerized marking of free-text responses.

[19]. **In M. Danson (Ed.),** Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughboroug University, Loughborouh, UK.

[20]. **Page, E.B. (1996).** Grading essay by computer: Why the controversy? Handout for NCME Invited Symposium.

[21]. **Page, E.B. (1994).** New computer grading of student prose, using modern concepts and software. Journal of Experimental Education, 62(2), 127-142.

[22]. **Palmer, J., Williams, R., & Dreher H. (2002).** Automated essay grading system applied to a first year university subject- How can we do it better. Proceedings of the Informing Science and IT Education (InSITE) Conference, Cork, Ireland, 1221-1229.

[23]. **Rudner, L.M. & Liang, T. (2002).** Automated essay scoring using Bayes' Theorem. The Journal of Technology, Learning and Assessment, 1(2), 3-21.

[24]. **Siddhartha Ghosh, Sameen S Fatima, (2007),** use of local languages in Indian portals, CSI Communication,

June'07 issue, pp- 4-12.

[25]. **Siddhartha Ghosh, Sameen S Fatima, (2007),** Retrieval of XML data to support NLP applications, ICAI'07- The 2007 International Conference on Artificial Intelligence Monte Carlo Resort, Las Vegas, Nevada, USA ,June 25-28, 2007.

[26]. **Siddhartha Ghosh, Sameen S Fatima, (2007),** A Web Based English to Bengali Text Converter, will be presented in The 3rd Indian International Conference on Artificial Intelligence (IICAI-07), Pune , India, December 17-19, 2007.

[27]. **Thompson, C. (2001).** Can computers understand the meaning of words? Maybe, in the new of latent semantic analysis. ROB Magazine. Retrieved from http://www.vector7.com/client_sites/ROB_preview/html/thompson.html

[28]. **Valenti, S., Cucchiarelli, A., & Panti M. (2000).** Web based assessment of student learning. In A. Aggarwal (Ed.), Web-based Learning & Teaching Technologies, Opportunities and Challenges, 175-197. Idea Group Publishing.

[29]. **Valenti, S., Cucchiarelli, A., & Panti, M. (2002).** Computer based assessment systems evaluation via the ISO9126 quality model. Journal of Information Technology Education, 1 (3), 157-175.

## ABOUT THE AUTHORS

*Associate Professor, Department of Computer Science & Engineering, G. Narayanamma Institute of Technology & Sc. Hyderabad , Andhra Pradesh, India.

**Associate Professor, BITS Pilani - Dubai Campus, UAE, on lien from Dept. of CSE, Osmania Universty, Hyderabad, Andhra Pradesh, India.

*Siddhartha Ghosh did his Completed B.Tech. from TEC, Agartala presently NIT Agartala, Tripura. He has an M.Tech. degree from University of Hyderabad. He is pursuing Ph.D. from Osmania University, Hyderabad. He has eight years of Teaching experience and two years Industry experience. He has 15 national and international publications to his credit. He is member of CSI, ISTE and IE. He has vast experience in NBA accreditation work and ISO 9000:2001 certifications for Educational institutes. His areas of interest are Artificial Intelligence, Natural Language Processing, Machine Learning, Web Technologies and Indian Languages.*

*S Sameen Fatima is a Professor in the Department of Computer Science and Engineering at Osmania University. She is currently on a sabbatical at BITS Pilani - Dubai. Her research interests include Information Retrieval Systems, Computational Linguistics and Text Mining. She received her B.Tech. from Jawaharlal Nehru Technological University in Electronics and Communication Engineering and got her M.Phil. in Computer Methods from Hyderabad University. She joined the Department of Computer Science and Engineering, Osmania University as a faculty member in 1984 during its formative years. During 1989-94, she received her M.S. in Computer Science and worked at the University of Massachusetts at Amherst, USA.*