Towards the use of Semi-structured Annotators for Automated Essay Grading

Hon Wai Lam (Sean), Prof. Tharam Dillon, Prof. Elizebeth Chang
Digital Ecosystems and Business Intelligence Institute
Curtin Business School
Curtin University of Technology
Perth, Australia
{S.Lam; Tharam.Dillon; Elizabeth.Chang}@cbs.curtin.edu.au

Abstract – The amount of time teachers spend grading essays has increased over the past decade, prompting the development of systems that are able to lighten the workload. Many systems have thus far used linear regression or semi-supervised methods towards this objective. This paper discusses some of the main Automated Essay Grading systems, highlighting some of their strengths and weaknesses, in addition to providing a brief overview of Text Mining and meta-data annotation techniques that could be used to facilitate the process of grading essays through an automated system.

Keywords- Automated Essay Grading; meta-data annotation; Text Mining; Named Entity Recognition; Part-of-speech Tagging

I. INTRODUCTION

The vast number of essays that teachers have to go through during marking has always been an issue. The task is relatively monotonous and costly; often taking several hours. Thus the development of an automated method to grade these essays was an important step. There have been many debates on the effectiveness of using a machine to grade an essay, the most common being that a machine would never have the same cognitive capabilities of a human reader and would thus be unable to give a score that considers more subtle aspects of written work. Discerning the implicit meaning of unstructured information such as that in the form of written text is also a problem faced by other fields in information retrieval and many solutions have been proposed under the general field of Text Mining. These techniques have been able to detect entities in text, but the task of linking them together and forming relationships to aid understanding and disambiguation is a more complex problem.

The field of automated essay grading can be considered relatively new and in its infancy, but it has a 40 year history [1]. One of the earlier programs designed for automated scoring was by Ellis Page in 1966, called the Project Essay Grader [2]. Using multiple linear regression, tried to pick out which weighted features of a text were most relevant to that of a grade given by a marker and in turn use those features to predict the score of an essay. Since then, there have been several other developments in essay scoring software that use a multitude of different techniques such as Latent Semantic Analysis, Natural Language Processing and Artificial Intelligence to name a few [3-6].

The remainder of this paper is organised as follows. In section II we discuss recent research into the fields of Automated Essay Grading systems culminating in a table comparing them. Section III presents some key concepts of this project which include techniques of Text Mining and Named Entity Recognition. Section IV describes the approach taken to automatically grading plain text essays and Section V concludes.

II. RELATED WORK

A. *Project Essay Grader (PEG)*

The Project Essay Grader developed in 1966 could be considered the pioneer of today's essay grading systems. The main idea behind Page's proposal is that it was possible to identify which features of a passage have the most influence on the score that a human rater would give; once those features are identified, multiple regression is used to compute a predictive formula of the score of an essay. 'Trins' (intrinsic) relate to the intrinsic aspects of an essay (e.g. fluency, diction, grammar, punctuation, etc) which Page determined to have a high weighting in the eyes of a human grader while 'Proxes' (approximated) refer to the correlation of those intrinsic variables [1, 2].

The scoring stage uses the two main variables, Trins and Proxes, gathered in the training stage from a test sample of 100-300 training essays to predict the score of an unmarked essay, the final score mainly depends on the linguistic aspects and style of an essay as evaluated by the PEG system [2, 6, 7]. An evaluation conducted by Page himself using roughly 30 Proxes found promising results, with the correlation between the PEG system and human graders to be .78 although this varies in later evaluations [7, 8]

A strength of the PEG system is the reasonably high correlation between human graders and the system generated score (some reaching as high as 0.85 between two or more graders); another is that the system is able to track errors, allowing for greater ease of evaluation [8, 9]. Having mentioned that, the weaknesses of the system are that since the contextual features of the essay such as organisation are not detected, constructive feedback is not given. Furthermore,

with only a surface scrutiny of the features, it is entirely possible to trick the system into giving a higher score by writing a longer essay with non-contextual reference to the topic [6, 8]. Since the 1990's PEG underwent some modifications in which several lexicons were added together with specific parsers.

B. Latent Semantic Analysis & Intelligent Essay Assessor

Foltz [10] described Latent Semantic Analysis (LSA) as a "statistical model of word usage that permits comparisons of the semantic similarity between pieces of textual information". In other words, LSA can be seen as a technique to analyse the relation between textual documents and their terms through a series of concepts. The whole idea behind LSA is that the meaning of a body of text is dependent on the meaning of each and every one of the words used and modifying any word would affect the meaning of the passage in one way or another [6]. As described by [3], "...meaning of word 1 + meaning of word 2 + ... + meaning of word n = meaning of passage ". This might lead to a possibility wherein passages that contain different words might have the same meaning and vice versa.

Developed by the University of Colorado and purchased by Pearson Knowledge Technologies (PKT), the Intelligent Essay Assessor (IEA) is one AEG system that uses LSA. The system places more emphasis on the context of the text rather than the common approach of scoring based on formal aspects such as grammar and punctuation although those are also incorporated into the scoring model [6, 7]. In this approach, a matrix is used to represent text, with each row representing a unique word while columns refer to the context in which it is used.

A prominent issue with the IEA is the number of essays required for training (roughly in the vicinity of 100-300), which even the producers of the system, PKT, state as a feature to improve on, although other systems have an even higher number required (upwards of 300). Furthermore, for all the analysis on content that the system performs, creativity as well as critical and reflective thinking by the student is not taken into account when calculating the essay score [3, 6].

One of the advantages stated by [8] and reiterated by [1] is that the system is able to "capture transitivity relations and collocation effects among vocabulary terms, thereby letting it accurately judge the semantic relatedness of two documents regardless of their vocabulary overlap" [6]. Above all, what makes IEA stand apart from other current systems is its ability to detect plagiarism, which escapes most human markers since it is a tedious task to perform, especially when a large number of essays are to be graded. A survey conducted by [7] reinforces the above mentioned points when 327 essays were sent for grading by IEA. The system managed to detect a few cases of plagiarism that had escaped the notice of human graders [7].

C. E-Rater & E-Rater V.2

Out of the few AEG systems that consider the linguistic features of a passage, E-Rater developed by the Educational Testing Service is another. This system incorporates Natural Language Processing (NLP) techniques which are used to pick out specific features from a test bed of sample essays which provide the basis of the scoring model [11]. The general assumptions or axioms of the e-rater system are that good essays would not be that much different from another good essay and likewise for poor essays.

The above mentioned features include a syntactic module, in which a parser is used to "identify [ies] syntactic structures, such as subjunctive auxiliary verbs and a variety of clausal structures, such as complement, infinitive, and subordinate clauses" [11] to pick out syntactic variety; a discourse module, in which a conceptual framework based on relations between conjunctions such as cue words (e.g. "probable" or "likely" to express a chance or probability), terms which could be in the form of conjuncts ("to summarise" or "to conclude" when summarising a passage) and syntactic structures, are used to consider the organisation and structure; finally, the topical analysis module picks out topical content and variety in the vocabulary [4, 11, 12]. Having identified the weighted features which make up a good essay, the e-rater then compares every new essay it evaluates against those features.

While the system as evaluated by Burstein and others in 1998 found the agreement rating between the system and human graders to be as high as 94% [11], the fact remains that the system does not actually go through or perform an actual analysis of the text since the scoring model is derived from the sample essays and every new essay is graded against it. Even though the system incorporates a set of more than 60 features [13]. Powers stated in an evaluation of the e-rater that it is not yet ready to function without human intervention, which is required to "keep e-rater from seriously misscoring some essays." [14]

The mechanics of the e-rater V.2 scoring system remains largely similar but improves on its predecessor by significantly reducing the amount of features by condensing them into a smaller set of more meaningful features which include Grammar, Style Measures, Organisation, Lexical Complexity and Prompt-specific Vocabulary Usage [13]. The other improvement from the first version is that it allows for a greater degree of standardisation by its ability to create a single scoring model from the feature set. The issues mentioned before are still present; while the feature module 'Lexical Complexity' considers word-based characteristics, key word frequency and word length do not necessarily measure the creativity of the writer per se.

D. IntelliMetric

Probably one of the first to utilize Artificial Intelligence (AI) into its scoring model, IntelliMetric developed by Vantage Learning in 1997 is used widely across the United States [5]. While many details of IntelliMetric remain a closely guarded secret by Vantage Learning, the general architecture uses a mixture of AI, Natural Language Processing (NLP) and Statistical tools, which create a sort of "neuro-synthetic" [6] logic processing which is said to mimic the way a human would think or "brain-based" according to [15].

Broadly speaking the final score of an essay assessed by the system is based on a set of five feature dimensions, namely [6, 15, 16].

- Focus and Unity – attributed to cohesiveness and consistency in the writers' focus on the main idea
- Development and elaboration – relates to the expansiveness of content and support for arguments
- Organization and Structure – measures discourse logic and transitional fluidity within the passage
- Sentence Structure – complexity of language use
- Mechanics and conventions –relates to adherence to standard English language rules

As found by an evaluation conducted by Rudner [5], the IntelliMetric system was able to closely match the scores as they were given by human graders, with the only small issue being that the system tended to give slightly higher scores, but a further investigation on the researchers' part concluded the issue possibly groundless seeing that both scores by human graders and the system fluctuated either way. Overall, the evaluation ended extremely favourably to the IntelliMetric system. Another attribute worth mentioning about IntelliMetric is its ability to evaluate essays of other languages, such Spanish, Hebrew, Dutch and French [17].

E. MarkIt

A more recent essay grading tool, MarkIt was proposed by [18] which made use of a rough clustering or "chunking" of the text in order to obtain sentence structure, represented by Noun Phrases and Verb Clauses which relate to the context and actions pertaining to the subject respectively.

According to the developers, Verb Phrases are extremely complex, thus prompting them to use Verb Clauses together with Noun Phrases. By mapping the root meaning of the word to the one found in the text, thereby assigning it the thesaurus index number, a numerical representation of the text can then be established. These are then used in a classification approach of predicting an essay's score using multiple linear regression, with vector space computations formulating some of the calculation inputs.

Some of the issues that can be identified with such a method are that the system seems to use nothing more than a version of Named Entity Recognition, where Noun and Verb Clauses are identified and counted, similar to the Bag of Words approach. While MarkIt might produce an accurate score with a high agreement rate among human graders under some circumstances, it would be easy to trick the system into giving a high grade if the mechanics of the algorithm are known even generally (e.g. including more keywords to attain a higher Noun Phrase value). Furthermore, it does not seem that the system is able to handle word sense disambiguation. A short in-depth evaluation conducted by [18] showed small inconsistencies with human graders and the IEA system, although there were some cases with larger differences.

F. General Comments

The general trend of most systems developed for automated scoring is to first perform a selection of certain salient features that make up a good or bad essay, then mapping those features onto an unmarked essay to attain a grade. This is usually done though linear regression or a semi-supervised machine learning method, in that training data (which varies between systems), is required for the system to learn either the model answer or a set of criteria on which the essay is graded on. Such methods are common in the field of Text Mining which is further discussed in the next section. Adding to the above, a table denoting a summary from previous works [6, 18-20] is given in table 1.

Although escaping mention in most literature on AEG systems, Text Mining Tools on most occasions provide the 'backbone' of these systems in that it is based on those tools that parts of speech or entity recognition is performed. The next section will briefly cover some of the key concepts of this project, namely Named Entity Recognition (NER), Part of Speech tagging (POS), and the annotation and mining of plain text.

AEG System	Developer	Year	Technique	Training Essays Required	Issues
Project Essay Grader (PEG)	Ellis Page	1996	Multiple Linear Regression	100-400	Relative large amount of training essays required
Intelligent Essay Assessor (IEA)	Landauer, Foltz and Laham	1998	Latent Semantic Analysis	300	Similar to PEG
E-rater	Jill Burstein	1998	Natural Language Processing techniques and Statistical computation	270-465	Prone to major errors without human supervision
IntelliMetric	Scott Elliot	1998	Artificial Intelligence and Statistical computation	300	Relative large amount of training essays required
MarkIt	Heinz Dreher and Robert Williams	2000	Multiple Linear Regression Vector space computation	50-100	Possible to trick system (e.g. having a high number of keywords)

Table 1: Comparison of AEG Systems

III. KEY CONCEPTS

A. Text Mining

There have been many ways that Text Mining has been described, a definition given by [21] says "Text mining techniques are dedicated to the automated information extraction from unstructured textual data". This opinion is also shared by other researchers [22, 23], who mention in one

way or the other that main aim of text mining is to discover knowledge from unstructured text, the word 'discover' thus implying that the information already exists and is just waiting to be found.

Bloehdorn [24] stated that currently, text mining techniques can be seen heading in two general directions. In the first, researchers try to conduct information retrieval and mining techniques through methods employing various degrees of automation. These methods, broadly speaking, fall under one of three different types of machine learning: supervised, semi-supervised and unsupervised learning. In some sub-tasks of Text Mining (e.g. Named-Entity Recognition), a common approach in supervised learning is where a set of annotated or labelled data that provides positive and negative examples is used to train the system; in the task of text categorization, categories that the documents or data are to be grouped into have already been predefined. Also, supervised learning involves a fair bit of human intervention.

Semi-supervised learning requires a lesser amount of human intervention, which can be in the form of manually annotating the training set or handcrafting a set of decision rules. Finally, unsupervised learning, in which certain criterion might or might not be determined beforehand to group unannotated documents into meaningful categories or to extract meaningful data from. Examples of sub-tasks of Text Mining that utilize unsupervised methods include document clustering and some NER/extraction techniques.

The second direction heads toward the fields of semantics and ontology, involving more metadata and conceptual structures to 'organize' textual information into a structured format. Most text mining techniques avoid deeper, more cognitive aspects of NLP and favour simpler techniques more like those used in practical information retrieval.

The main difficulty in text mining is not that the information is hidden in textual documents but that the data that resides in these documents is unsuitable for computer processing, setting it slightly apart from data mining, in which the data is already in some form of structured format, although the two share many of the same techniques; e.g. categorization, clustering, decision trees.

B. Annotation of Plain Text

Much work has been done in the aims of managing unstructured information mostly in the form of natural language text. In this project, text is analysed and annotated with meaningful tags that allow for a more structured reading. Through this step, it is possible to discern meaning from free text otherwise hidden to a machine reading component. One such method of organising unstructured information is IBM Research's unstructured information management architecture, which is further described below.

1) UIMA

IBM Research's unstructured information management architecture (UIMA) is mainly focused on developing components that can be implemented in a number of ways for the purposes Natural Language Processing. According to [25], they characterize UIMA as "a software system that analyses large volumes of unstructured information". One of the advantages of UIMA is that components that make up the architecture, such as the Text Analysis Engines (TAE), are highly portable in the sense that they can be developed by different software engineers, packaged and reused in another setting. TAE's usually perform natural language tasks such as stemming or NER, which then pass that information onto other analysis engines or to the end user components.

C. Named Entity Recognition

Downey [26] defined the process of Named Entity Recognition (NER) as the task of identifying and classifying names in textual documents. An alternative description is where NER is a subtask of Information Extraction in which string elements are grouped into predefined categories such as persons, organisations or locations. In a more generalised explanation, Alfonseca and Manandhar [27] state that NER involves the identification and classification of instances or objects of interest, which can fall under the above categories or "anything that is useful to solve a particular problem".

In order to effectively and correctly extract information Text Mining tools need to be able to distinguish which words or 'linguistic constructions', represent entities [28]. Early NER tools used a set of rules that were input manually, which much like the problems faced in Brute Force type algorithms, take too much effort to correct and maintain. Modern methods of extracting entities are more inclined towards, though not limited to, the use of supervised methods in which an NER tool is first trained on a limited number of documents and by using one of several machine learning techniques, enabling the tool to automatically decide which strings elements constitute an entity.

Entities are usually represented by more than one word but are seen as single vocabulary strings by NER tools (e.g. the name 'Jane Smith' or the company 'General Motors'). For example, consider the sentence, "Nokia was founded by Fredrik Idestam in Finland". 3 named entities are present: 'Nokia' which is an organization, 'Fredrik Idestam' is a person and 'Finland' is a location. The entities described above are the most commonly extracted by NER tools, generally termed as 'proper names' [29].

D. Part of Speech Tagging

The main aim of POS is to assist in recognizing patterns in natural language documents by automatically assigning tags (nouns, adjectives, verbs, etc) to words in the document's context, which facilitates more advanced analysis techniques in the text mining scope. Difficulties commonly faced in part of speech tagging are the lexical ambiguities that exist on most natural language documents. For instance, the words 'process' and 'programs' could be both tagged as verbs or nouns, although this problem can be partially bypassed through the analysis of the context of the text itself [30].

Most part of speech tagging is primarily split into 2 steps, in the first term "candidates" are extracted based on the structure of the linguistic information, in other words the context of the text. For example, candidates can be selected

based on morpho-syntactic patterns such as noun-noun (George Clooney) or noun-preposition-noun (Head of State). Those candidates are then filtered according to one or more of a type of statistical relevance scoring scheme such as frequency of occurrence, similar information, log-like coefficients, etc. [31].

IV. Methodology/Approach

The first step in our approach is to be able to tokenise unstructured text so as to facilitate machine reading and analysis. This is done with the aforementioned UIMA; using a simple annotator, it is possible to tokenise each word in a paragraph, in turn allowing it to be parsed into different parts of speech or named entities. A sample output generated using a simple token annotator is given below:

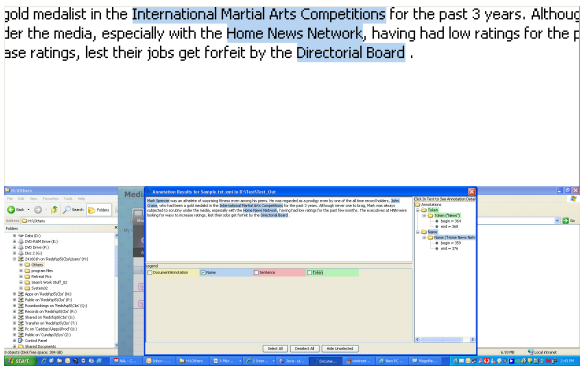

Figure 1: Sample output from token annotator (top half is an enlarged view of the text)

The information can also be presented in an XML format, which further facilitates analysis of the text, a small sample output of the same text in XML form is given below:

```
org.apache.uima.examples.tokenizer.Token
sofa="Sofa"                                    begin="359"
end="363">Home</org.apache.uima.examples.token
izer.Token>
<org.apache.uima.examples.tokenizer.Token
sofa="Sofa"                                    begin="364"
end="368">News</org.apache.uima.examples.tokeni
zer.Token>
<org.apache.uima.examples.tokenizer.Token
sofa="Sofa"                                    begin="369"
end="376">Network</org.apache.uima.examples.tok
enizer.Token>
        </example.Name>
```

Figure 2: Sample output in XML

Using the annotator, it is possible to tokenise named entities such as "Home News Network". The main idea of the next stage is thus representing these annotations within a structure which would allow for the mining of more meaningful information. The full text with the named entities is shown below:

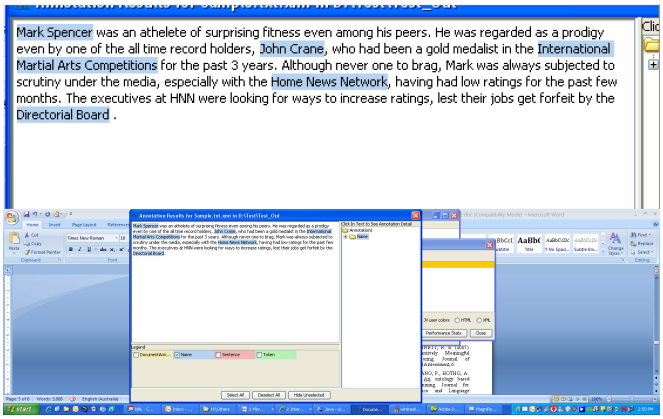

Figure 3: Full text with Named Entities (top half is an enlarged view of the text)

From the above, the key points in the short essay can be easily determined. It can be seen that there are six entities in this passage and that much of the events happen around them. In addition, it is also necessary to identify the tokens in as parts of speech to further the analysis process. The part of speeches of the full text is shown below:

Word	*POS Tag*	*Word*	*POS Tag*	*Word*	*POS Tag*	*Word*	*POS Tag*
Mark	NP	John	NP	always	RB	were	VBD
Spencer	NP	Crane	NP	subjected	VVN	looking	VVG
was	VBD	,	SENT	to	TO	for	IN
an	DT	who	WP	scrutiny	NN	ways	NNS
athelete	NN	had	VHD	under	IN	to	TO
of	IN	been	VBN	the	DT	increase	VV
surprising	JJ	a	DT	media	NNS	ratings	NNS
fitness	NN	gold	JJ	,	SENT	,	SENT
even	RB	medalist	NN	especially	RB	lest	IN
among	IN	in	IN	with	IN	their	PP$
his	PP$	the	DT	the	DT	jobs	NNS
peers	NNS	ternationa	NP	Home	NP	get	VVP
.	SENT	Martial	JJ	News	NP	forfeit	VV
He	PP	Arts	NP	Network	NP	by	IN
was	VBD	ompetitio	NNS	,	SENT	the	DT
regarded	VVN	for	IN	having	VHG	Directorial	NP
as	IN	the	DT	had	VHD	Board	NP
a	DT	past	JJ	low	JJ	.	SENT
prodigy	NN	3	CD	ratings	NNS		
even	RB	years	NNS	for	IN		
by	IN	.	SENT	the	DT		
one	CD	Although	IN	past	JJ		
of	IN	never	RB	few	JJ		
the	DT	one	CD	months	NNS		
all	DT	to	TO	.	SENT		
time	NN	brag	VV	The	DT		
record	NN	,	SENT	executives	NNS		
holders	NNS	Mark	NP	at	IN		
,	SENT	was	VBD	HNN	NP		

Table 2: Part of Speech Tags

With the above information, coupled with a marking rubric, it is thus possible to determine the extent of which this essay pertains to a given topic. In order for this project to have a meaningful use, the scope is currently set within the WA marking rubric as stated in the National Assessment Program.

V. Conclusion & Future Directions

This paper has presented some of the main essay grading systems currently available, including a discussion of some of the main issues relating to those systems. A short over view of Text mining and meta-data annotation techniques was also provided, highlighting key concepts and the potential of using these techniques to facilitate the process of grading essays through an automated process. The use of these meta-data annotation techniques for a small portion of text similar to an essay was demonstrated.

The next step in this project would detail how the above annotations are used in the analysis of unstructured text. This analysis will require the use of an ontology based on the rubric as well as an ontology for the domain of interest. This process thus allows the characterisation of both structure and semantics of the essay, which are essential in carrying out in-depth automated essay grading.

VI. References

[1] J. Wang and M. S. Brown, "Automated Essay Scoring Versus Human Scoring: A Comparative Study," *Journal of Technology, Learning and Assessment,* vol. 6, 2007.

[2] E. B. Page, "The imminence of grading essays by computers," *Phi Delta Kappan,* vol. 47, pp. 238-243, 1966.

[3] T. K. Landauer, D. Laham, and P. W. Foltz, "Automated Essay Scoring and annotation of essays with the Intelligent Essay Assessor," in *Automated Essay Scoring: a cross-disciplinary perspective*, M. a. B. J. C. Shermis, Ed.: Lawrence Erlbaum, 2003, pp. 87-112.

[4] J. Burstein, M. Chodorow, and C. Leacock, "Criterion: Online essay evaluation: an application for automated evaluation of student essays," in *Proceedings of the 15th Annual Conference on Innovative Applications of Artificial Intelligence*, Acapulco, Mexico, 2003.

[5] L. M. Rudner, V. Garcia, and C. Welch, "An Evaluation of the IntelliMetric," *Journal of Technology, Learning and Assessment,* vol. 4, 2006.

[6] S. Dikli, "An overview of Automated Scoring of Essays," *Journal of Technology, Learning and Assessment,* vol. 5, 2006.

[7] R. Williams, "Automated essay grading: an evaluation of four conceptual models," in *10th Annual Teaching Learning Forum* Perth, Curtin University of Technology: Herrman A and Kulski M, Expanding Horizons in Teaching and Learning, 2001.

[8] K. Kukich, "Beyond automated essay scoring," *IEEE intelligent systems,* vol. 15, p. 22, 2000.

[9] G. K. W. K. Chung and J. H. F. O'Neil, "Methodological Approaches to Online Scoring of Essays," University of California 1997.

[10] P. W. Foltz, "Latent Semantic Analysis for text-based research," *Behavior Research Methods, Instruments and Computers,* vol. 28, pp. 197-202, 1996.

[11] J. Burstein, "The e-rater scoring engine: Automated Essay Scoring with natural language processing," in *Automated Essay Scoring: A cross disciplinary approach*, M. D. S. a. J. C. Burstein, Ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2003, pp. 113-121.

[12] J. Burstein and D. Marcu, "Benefits of modularity in an Automated Essay Scoring System," Education Resources Information Center, ED No. 447168, 2000.

[13] Y. Attali and J. Burstein, "Automated Essay Scoring with e-rater V.2," *Journal of Technology, Learning and Assessment,* vol. 4, 2006.

[14] D. E. Powers, J. Burstein, M. Chodorow, M. E. Fowles, and K. Kukich, "Stumping e-rater: challenging the validity of automated essay scoring," *Computers in Human Behavior,* vol. 18, pp. 103-134, 2002.

[15] S. Elliot and C. Mikulas, "The impact of MY Access! use on student writing performance: A technology overview and four studies from across the nation. ," in *Annual Meeting of the National Council on Measurement on Education, April 12-16* San Diego, CA, 2004.

[16] A. Ben-Simon and R. E. Bennett, "Toward More Substantively Meaningful Automated Essay Scoring," *Journal of Technology, Learning, and Assessment,* vol. 6, 2007.

[17] S. Elliot, "IntelliMetric: from here to validity. ," in *Automated essay scoring: A cross disciplinary approach*, M. D. S. a. J. C. Burstein, Ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2003.

[18] R. Williams and H. Dreher, "Automatically Grading Essays with Markit," *Issues in Informing Science and Information Technology,* 2003.

[19] M. Warschauer and P. Ware, "Automated writing evaluation: defining the classroom research agenda," *Language Teaching Research,* vol. 10, pp. 157-180, 2006.

[20] T. Kakkonen, N. Myller, E. Sutinen, and J. Timonen, "Comparison of Dimension Reduction Methods for Automated Essay Grading," *Educational Technology and Society,* vol. 11, pp. 275-288, 2008.

[21] M. Rajman and R. Besancon, "Text Mining: Natural Language Techniques and Text Mining Applications," in *7th IFIP 2.6 Working Conference on Database Semantics* Leysin, 1997.

[22] R. Feldman, Y. Aumann, M. Fresko, O. Liphstat, B. Rosenfeld, and Y. Schler, "Text Mining via Information Extraction " in *Principles of Data Mining and Knowledge Discovery*. vol. 1704/1999: Springer Berlin, Heidelberg, 1999.

[23] Y. N. Un and R. J. Mooney, "Text mining with information extraction," *American association for artificial intelligence,* 2002.

[24] S. Bloehdorn, P. Cimiano, A. Hotho, and S. Staab, "An ontology based framework for text mining," *Journal for Computational Linguistics and Language Technology,* vol. 20, pp. 87-112, 2004.

[25] D. Ferrucci and A. Lally, "UIMA: an architectural approach to unstructured information processing in the coporate research environment," in *Natural Language Engineering*. vol. 10: Cambridge University Press, 2004, pp. 327-348.

[26] D. Downey, M. Broadhead, and O. Etzioni, "Locating Complex Named Entities in Web Text," in *Proceedings of IJCAI*, 2005.

[27] E. Alfonseca and S. Manandhar, "An unsupervised method for general named entity recognition and automated concept discovery " in *Proceedings of the International Conference on General WordNet*, 2002.

[28] I. H. Witten, *Text mining*: CRC Press, 2004.

[29] D. Nadeau, P. Turney, and S. Matwin, "Unsupervised Named Entity Recognition: Generating Gazetteers and Resolving Ambiguity," in *19th Conference on Artificial Intelligence.* Canada, 2006.

[30] D. Cutting, J. Kupiec, j. Pedersen, and P. Sibun, "A practical part of speech tagger," in *3rd Conference on Applied Natural Processing*, ACL, 1992.

[31] B. Daille, "Study and implementation of combined techniques for automated extraction of terminology," *The Balancing Act: Combining Symbolic and Statistical Approaches to Language. MIT Press, Cambridge,* 1994.