Using Learning Analytics to Analyze Writing Skills of Students: A Case Study in a Technological Common Core Curriculum Course

Chi-Un Lei, Ka Lok Man and T.O. Ting

Abstract—**Pedagogy with learning analytics is shown to facilitate the teaching-learning process through analyzing student's behaviours. In this paper, we explored the possibility of using learning analytics tools Coh-Metrix and Lightside for analyzing and improving writing skills of students in a technological common core curriculum course. In this study, we i) investigated linguistic characteristics of student's essays, and ii) applied a machine learning algorithm for giving instant sketch feedback to students. Results illustrated the necessity of improving student's writing skills in their university learning through e-learning technologies, so that students can effectively circulate their ideas to the public in the future.**

Index Terms—**General education, learning analytics, educational data mining, computational linguistics, text analysis, automated essay scoring**

I. Introduction

With the recent advances in information technologies, an emerging mode of practices known as the learning analytics (educational data mining) has begun to change the paradigm of higher education [1]–[3]. Learning analytics can be defined as "*the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs*". It has been used to model individual learning contents and trajectories as well as social learning behaviours.

Meanwhile, writing is a major class of discourses and evidences that can give us insights into deeper learning and high-order skills such as critical thinking, argumentation and mastery of complex ideas. However, evaluating essays is an effort-demanding task, and usually teachers can only provide limited feedback or guidance to students, throughout the student writing process. Thus, various computational linguistics [4] and automated essay scoring (AES) techniques [5]–[7] have been adopted for teaching enhancements.

Currently, Hong Kong is adopting the higher-education transformation from a three-year curriculum to a four-year curriculum. With that in mind, the University of Hong Kong (HKU) has adopted new technologies and practices for teaching facilitations [8]–[11]. In particular, in order to provide key common learning experience for all undergraduate students and to broaden their horizons beyond their chosen disciplinary fields of study, HKU has introduced Common Core Curriculum (CCC). One of the CCC goals in HKU is to cultivate students to play an active role as responsible individuals in communities. Thus, in order to help students circulate their ideas to the public effectively in the future, CCC is also responsible to help students with their writing process. However, due to the tight teaching schedule, it is difficult for instructors to instantly examine student's writing. Therefore, we would like to explore the feasibilities of using learning analytics for developing student's writing skills.

In this paper, we give a discussion on applying textual learning analytic tools for analyzing and improving student's writing skills. Contributions of our paper are as follows:

- We have used a computational linguistic tool Coh-Metrix [12] to analyze the readability and linguistic features of student's essays. Once these features are identified, we can help students overcome the obstacles that less cohesive texts might present.
- We have used an AES system Lightside [13] for preliminary essay marking. We hope that the tool eventually can be used for self-directed learning of writing.
- We have analyzed the language varieties and discourse characteristics of writings for differences between three-year curriculum students and four-year curriculum students under the education reformation.

Section II describes linguistic analysis procedure through Coh-Metrix. Meanwhile, Section III describes the AES procedure through Lightside.

II. Linguistic Analysis via Coh-Metrix

Computational linguistics study languages from a computational perspective. Through knowledge-based or data-driven modeling, linguistic phenomena and behaviours can be modeled by computational models. These models are often used as a working component of a language system. For example, readability has been generally described by the Flesch Reading Ease (FRE) Score

$$\mathrm{FRE} = 206.835 - (1.015 \times \mathrm{ASL}) - (84.6 \times \mathrm{SW}), \quad (1)$$

where ASL is the average sentence length or the number of words divided by the number of sentences; SW is the average number of syllables per word. A higher FRE score indicates that the article is easier to read. Generally, an essay should have a Flesch Reading Ease score between 6 and 70.

Meanwhile, in this paper, we mainly focused on a few sophisticated linguistic characteristics of texts: syntactic simplicity, word concreteness, referential cohesion, and deep cohesion. They mainly describe whether the essay is helping the

Manuscript received June 6, 2014.

C.-U. Lei is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong. Email: culei@eee.hku.hk

K.L. Man is with the Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, China and Baltic Institute of Advanced Technologies, Lithuania. Email: ka.man@xjtlu.edu.cn

T.O. Ting is with the Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, China. Email: toting@xjtlu.edu.cn

(Advance online publication: 23 August 2014)

reader mentally connect ideas in the text and whether is easy to comprehend. These characteristics have been discussed in [12], and are outlined in the following subsection.

A. Studied Course and Essays

The studied course, Everyday Computing and the Internet (CCST9003) is a CCC course first offered in 2010. Besides introducing students a "computational thinking" concept through twelve-weeks teaching, CCST9003 also discusses intensively the societal impacts of computing technologies on our daily life, through surveying of computational methods and analyzing usage of computational methods.

In order to learn how to circulate ideas about computational thinking to the public, students have to write a survey essay on a topic related to everyday computing and the internet. The essay should offer knowledge and inspiration to the public as well as engagement with ideas.

B. Linguistic Characteristics of Essays

In this paper, each of these characteristics for a given text has been normalized, according to thousands of text samples stored in the Coh-Metrix database.

1) Syntactic Simplicity: Syntactic simplicity reflects the degree to which the sentences in the text contain fewer words and use simpler, familiar syntactic structures. Coh-Metrix measures syntactic simplicity through several indices. For example, texts with fewer clauses and words per sentence, and fewer words before the main verb/clause will give a text a higher score for syntactic simplicity.

2) Word Concreteness: Concrete words (e.g. apple, bottle, car and dog) are words that stimulate sensory response in the reader. In other words, we can imaginatively use our senses to experience what the words represent. On the other hand, abstract words (e.g. love, success, freedom and joy) usually refer to ideas or concepts with no physical referents. Coh-Metrix can compute the average Word Concreteness through a rating database of 4293 unique words. For example, words "protocol" (264) and "difference" (270) are recorded as less concrete than "ball" (615) in the database. A text with relatively high numbers of concrete words is easier to read, thus will have a high word concreteness score.

3) Referential Cohesion: A text with high referential cohesion contains words and ideas that overlap across sentences and the entire text, forming explicit threads that connect the text for the reader. When sentences and paragraphs have similar words or conceptual ideas (i.e. high referential cohesion), it is easier for readers to deduce connections between those ideas as well as to understand the essay. Referential cohesion can be measured by the overlap between verb, noun, argument, word stem and content word from one sentence to the other.

4) Deep Cohesion: Deep cohesion measures how well the events, ideas and information of the whole text are tied together. This can be measured by connectives and types of words that connect different parts of a text. For example, adversative connectives are words that connect two phrases or notions that conflict with each other, such as "*My favourite subject is operational management however I studied engineering.*" or "*Tomato is a fruit, yet it is used in savoury.*" These connectives are shown in Table I.

TABLE I
TYPES OF THESE CONNECTIVES.

Type	Connectives
Time	after, earlier, before, during, while, later
Causal	because, consequently, thus
Additive	both, additionally, furthermore, moreover
Logical	actually, as a result, due to
Adversative	but, yet, however, although, nevertheless

5) Connectivity: Connectives play an important role in the creating cohesive links between ideas and clauses and providing clues about text organization. Connectivity reflects the degree to which the text contains explicit connectives to express relations in the text. This component reflects the number of logical relations in the text that are explicitly conveyed. This score is likely to be related to the readers deeper understanding of the relations in the text.

6) Temporality: Texts that contain more cues about temporality and that have more consistent temporality (e.g., tense, aspect) are easier to comprehend. In addition, temporal cohesion helps readers understand the situation of the event in the text.

7) Length of Sentences and Paragraphs: The organization of an essay can also be described by the structure of sentences and paragraphs, the mean number of sentences in paragraphs and the mean number of words in sentences. A higher value indicates that the section may have more complex syntax and thus may be more difficult to process. For example, a large standard deviation of the mean length of paragraphs indicates that the essay may contain short and long paragraphs, posing understanding difficulty for readers.

C. Results and Discussions

We studied 25 essays from the three-year curriculum students and 26 essays from the four-year curriculum students. Furthermore, effect size has been calculated to show the strength of the relationship between variables. Results are shown in Table II. These metrics can assess students whether they can write organized and rich essays that are easy to understand. Thus, these metrics can be used to investigate problems in student's writing, for example,

- A low concreteness score indicates students may not be able to explain abstract ideas clearly.
- A low referential cohesion score means students might have trouble on building sentences on each other.
- A low deep cohesion scores indicates students have difficulties to comprehend how the ideas, events or information of the text as a whole fit together.

Based on Table II, some observations are as follows:

- Four-year curriculum students tend to write more paragraphs with less sentences and words in each paragraph. Usually it is not easy to develop a concrete idea in a paragraph with three or four sentences only, due to the lack of supporting details. For example, their essays tend to have less content overlap in terms of argument and content word. Thus, ideas developed by students may not be effectively circulated to the public.
- Essays written by four-year curriculum students tend to possess more syntactic simplicity and temporality, comparing to those from three-year curriculum students.

TABLE II
RESULTS OF ANALYSIS. (SD: STANDARD DEVIATION)

Metric	4-Year (Mean)	4-Year (SD)	3-Year (Mean)	3-Year (SD)	Effect Size
Flesch Reading Ease	50.71	7.06	**50.76**	7.74	-0.01
Number of paragraphs	16.6	8.26	10.46	7.38	0.78
Number of sentences	49.4	16.37	38.38	12.34	0.76
Number of sentences in a paragraph (Mean)	3.78	3.63	4.56	1.88	-0.27
Number of sentences in a paragraph (Standard Deviation)	3.69	6.42	2.36	1.10	0.29
Number of words in a sentence (Mean)	14.88	2.49	17.11	2.58	-0.88
Number of words in a sentence (Standard Deviation)	9.40	2.41	9.43	2.42	-0.01
Syntactic simplicity (Percentile)	**68.65**	13.34	60.94	19.30	0.46
Word concreteness (Percentile)	**25.64**	17.22	25.47	22.46	0.01
Referential cohesion (Percentile)	29.18	20.50	**33.62**	23.91	-0.20
Deep cohesion (Percentile)	66.92	18.54	**72.29**	19.31	-0.28
Verb cohesion (Percentile)	19.53	18.70	**39.57**	25.58	-0.89
Connectivity (Percentile)	**6.42**	9.48	5.93	13.85	0.04
Temporality (Percentile)	**45.28**	22.96	40.62	22.25	0.21
Noun overlap (Adjacent sentences)	0.38	0.13	**0.39**	0.12	-0.11
Argument overlap (Adjacent sentences)	0.46	0.13	**0.48**	0.13	-0.16
Stem overlap (Adjacent sentences)	0.49	0.13	**0.50**	0.09	-0.10
Noun overlap (Adjacent sentences)	**0.30**	0.12	0.29	0.12	0.03
Argument overlap (Adjacent sentences)	0.37	0.11	**0.37**	0.13	-0.04
Stem overlap (Adjacent sentences)	**0.40**	0.11	0.38	0.10	0.19
Content word overlap (Adjacent sentences)	0.09	0.03	**0.09**	0.04	-0.26
Content word overlap (All sentences)	0.06	0.02	**0.07**	0.03	-0.26

Meanwhile, three-year curriculum students work better in developing referential/deep/verb cohesion relationships in their essays. This may be due to lessening of writing training in the new high school curriculum.

- Comparing to samples written by students all around the world, students in HKU tend to write essays with a low overall cohesion and connectivity. This indicates their essays are less-organized and less easy to be understood. Thus, student's writing skills should be improved through appropriate instructions.

III. AUTOMATED ESSAY SCORING VIA LIGHTSIDE

Automated essay scoring (AES) [5]–[7] study essays and assign grades to essays written in an educational setting. In other words, through knowledge-based or data-driven modeling, essay contents and characteristics can be modeled quantitatively. These models can be used to classify a large set of textual cases into a number of discrete categories (i.e. grades). AES becomes popular recently because it can measure accountable educational achievement at reduced cost. Various AES systems, such as LightSide [13], have been proposed [5], [14].

Different from other AES systems, LightSide has been proposed for self-directed learning. To be specific, Lightside is not just checking on grammatical errors but also assessing essay contents by comparison of essays stored in the database. Through submitting their draft essay to the system, students can collect feedback on their drafts before the official submission. By iteratively assessing and revising, students are able to write an essay with better organizations and contents, as well as develop literacy skills and meta-cognitive skills in the iterative revision process.

A. General Procedure for AES

Adequate samples are needed for training before the system can mark essays by itself. Usually, the training file contains records of training cases. Each case contains the essay text, meta-data and a user-defined grade (i.e. the label). Labels can be nominal (discretized and limited) or numeric (real number). **Feature extraction** is used to generate a feature table from essay records. After the features table has been constructed, a **machine learning** process is used to discover the latent pattern in those features (similar to other information retrieval applications [15]) and train the model/classifier. The constructed model can then be validated and analyzed by instructors. After the **model validation**, the validated model can be used for **grade predication**.

B. Features Used for AES

Features describe the text, and can affect the properties and quality of the model. However, essay data are often noisy, thus meaningless features can be extracted if the extraction is not supervised. Therefore, different techniques and feature set have been proposed to ensure extracted features are maximally informative for classifications. Examples of popular feature set for AES are shown in the following subsections.

1) "Bag-of-Words" (BOW): In the BOW model, the text is represented as a bag of its words, disregarding grammar and word order. It is commonly used for document classifications, where the occurrence of each word is used as a feature for training. Examples of BOW are as follows:

- Unigram: Single word. An example is "Internet".
- Bigrams: Two consecutive words in a certain sequence. For example, bigram "the to" is different from "to the".
- Stem N-Grams: Words that are constructed from the basic form. For example, "walk", "walks" and "walking" can be grouped to a generic gram "walk".
- Stretchy Patterns/"N-Grams with gaps": Phase features with a small variations.

2) "Part-of-Speech" (POS): In the POS model (a.k.a. word class, lexical class, or lexical category) model, the text is represented as a bag of linguistic categories of words. These categories is generally defined by the syntactic behaviour of the lexical item (e.g. noun, verb). For example, the sentence "We are young." can be decomposed into the following POS:

- BOL_PRP: The beginning of a line followed by a personal pronoun.

- PRP_VBP: A personal pronoun followed by a non third-person singular present verb.
- VBP_JJ: That same verb part-of-speech tag, followed by an adjective.
- JJ_EOL: An adjective followed by the end of line.

POS is also commonly used in for document classifications.

3) Other Features and Feature Processing: Examples are class, curriculum, age, gender and user-defined text patterns. The system also allows extraction of parse features.

After extraction, extracted features can then be explored. Statistics such as total hits, target hits, precision (fraction of relevant instances that are retrieved), Kappa (how well it performed above chance) and correlation, is shown. Based on the statistics, meaningless features can be deleted and coupled features can be combined with logic compositions (e.g. "Internet OR Network"). Through feature processing, features become more meaningful for classifications.

C. Machine Learning

Machine learning algorithms are used to formulate a set of rules, based on training examples. Formulated rules are used for labelling or tagging documents with similar contents in the future. The following algorithms are often used for learning BOW feature spaces:

- Naive Bayes: It learns from each features individually, but not from the dependencies between features. It has been widely used for email spam filtering and other text classification situations.
- Logistic Regression: It is a common probabilistic statistical classification model for text mining.
- Support Vector Machine (SVM): It focuses on classifying marginal instances. Therefore, it is good at binary classification, but behaves poorly for cases with many possible labels.

After the machine training process, cross validation is usually needed to check whether the model behaves satisfactorily. Cross validation is to slice up the training data into "folds" (subsets), and hold out one fold each turn. For example, in four-fold cross validation, the training set is decomposed into four subsets. Then, in the first round, the first 3 subsets are used for training and the last subset for testing. In the second round, subsets 1, 2 and 4 are used for training and subset 3 for testing. By continuing the process a few times, a set of guesses at performance can be obtained.

The model performance can be shown by accuracy (how many examples it labelled correctly) and Kappa and a confusion matrix. Confusion matrix is a table with rows and columns that reports the number of documents with its corresponding grades marked by instructors and the machine. The intersection of identical row and column labels show the number of documents that the model has predicted the document's label correctly. On the other hand, other cells represent incorrect predictions.

D. Post-Learning Error Analysis

Before applying the model in a real-world classification, it is better to understand the behavior of the model, and calibrate settings of feature extraction and machine learning, for a better classification. In particular, features that significantly affect the classification and appear frequently should be checked. Lightside provides the following indicators:

- Frequency: The number of documents that contain the selected feature
- Average value: The average value that the selected feature has in documents
- Influence: The indicator that intuitively shows how the feature is associated with a particular prediction label

E. Studied Essays and Results

The studied course, Electronic Technologies in Everyday Life (CCST9015) is a CCC course first offered in 2010. Besides introducing to students knowledge of modern electronic technologies, CCST9015 also discusses the societal impacts of these technologies on our daily life.

After twelve-weeks teaching, students have to write a short sketch on a topic *"What do you know about electronic technologies in everyday life?"* as a learning consolidation. Sample essays have been classified by instructors into three categories/grades ("Good", "Fair", "Poor"). At the end of the course, we have collected 96 records (student's answer), and each record contains two main features: the student sketch and an instructor-assigned grade. Examples of student's essay and the corresponding grade are shown as follows:

- "Good" (I): *"Electronic technologies are everywhere and are an integral part of our everyday lives. We benefit a lot from electronic technologies from the speed of electronic gadgets and so on. However, they also pose threats such as possible addiction to usage of iPhone, possible electronic hypes that lead to unnecessary waste of resources, or even moral breakdown due to potential security breakdown such as loss of privacy."*
- "Good" (II): *"Electronic technologies are closed to our daily life because many tools using in our life are based on the electronic technologies. For example, the screen touch technology and the octopus card. Without these kind of things, our life become very inconvenient."*
- "Fair" (I): *"Electronic technologies have already integrated into everyday life, which we cannot live in the same life quality without them. One particular example would be computers and internet that enable online communication throughout the world."*
- "Fair" (II): *"The basic concept of the wireless network of the cellular phone, the security concern of the networking, the hardware of the electronic device (such as the CPU, battery etc.)"*
- "Poor" (I): *"LTE, wifi, processor, CPU, GPU"*
- "Poor" (II): *"Nowadays, we cannot live without electronic technologies. For example, nearly all the world is connected by the internet."*

Machine learning with a simple unigrams feature is first used for illustrations. 138 features have been extracted from these records. Based on the obtained features, Naive Bayes algorithm with ten-fold validations has been used for machine learning. Its model evaluation metric and confusion matrix are shown in Table III. Among 96 tested cases, 74 of them have been graded correctly by the machine. In particular, no cases marked as "Poor" by instructors have been marked as "Good" by the machine (and vice versa).

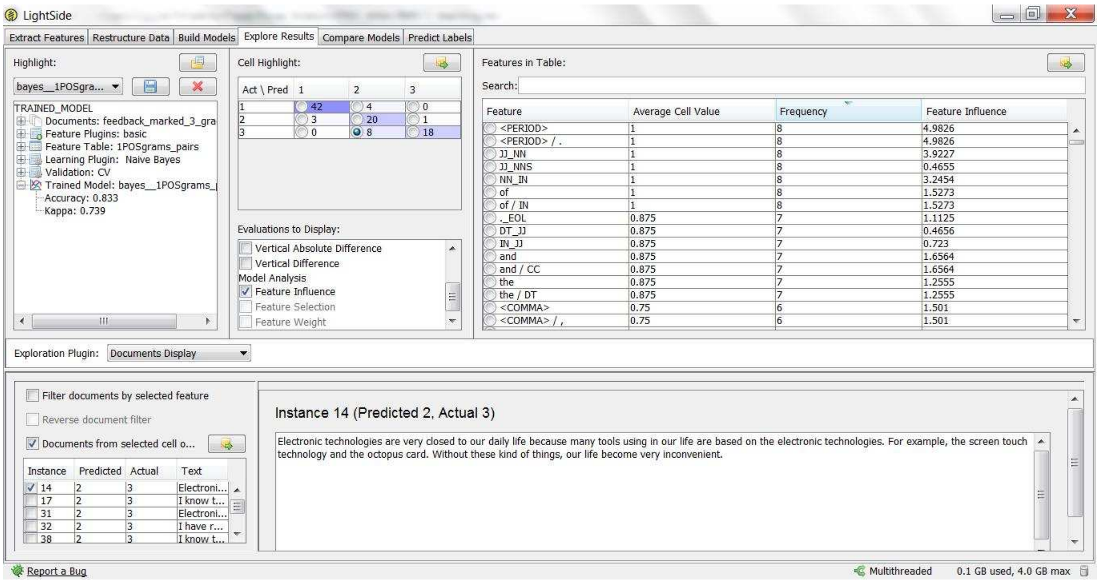

Fig. 1. Error analysis in the Lightside system.

TABLE III
MODEL EVALUATION METRICS AND CONFUSION MATRICES

Single-Feature Learning — Accuracy = 0.7708; Kappa = 0.6391			
Actual \ Predicated	Poor	Fair	Good
Poor	42	4	0
Fair	4	14	6
Good	0	8	18
Multi-Feature Learning — Accuracy = 0.8333; Kappa = 0.7390			
Actual \ Predicated	Poor	Fair	Good
Poor	42	4	0
Fair	3	20	1
Good	0	8	18

In the second example, a sophisticated feature set has been used for learning. The set consists of unigrams, POS bigrams, POS trigrams, word/POS pairs, stem N-grams and stretchy patterns. 961 features have been extracted from these records. Its model evaluation metrics and confusion matrix are shown in Table III. Among 96 tested cases, 80 of them have been graded correctly by the machine, as shown in Fig. 1. In particular, the performance of classifying essays with "Fair" grade is better (i.e. less essays have been misclassified as "Poor" and "Good"), compared to the previous example. This is because essays with "Good" grade and "Fair" grade have similar contents, but the former one usually is better organized. The discrepancy can be modeled by POSs and stretchy patterns but not unigram. With a comprehensive feature set, the machine learning process can discover more latent patterns and train a better model for classifications.

IV. CONCLUSION

In this paper, we have used Coh-Metrix and Lightside to analyze organizations and contents of students' essay in a technological Common Core Curriculum course. The evaluation illustrated the necessity of improving student's writing skills in their university learning stage. The quantitative analysis methodology can be extended to the determination of several advanced metrics, namely: lexical diversity, syntactic complexity, and syntactic pattern density. The machine learning methodology can be extended by introducing more sophisticated features that can directly measure (meta-)cogitative abilities and writing skills, in order to comprehensively assess abilities of students.

REFERENCES

[1] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," *Educause Review*, vol. 46, no. 5, pp. 30–32, 2011.
[2] R. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.
[3] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 40, no. 6, pp. 601–618, 2010.
[4] R. Ferguson and S. B. Shum, "Learning analytics to identify exploratory dialogue within synchronous text chat," in *Proc. ACM Intl. Conf. on Learning Analytics and Knowledge*, 2011, pp. 99–103.
[5] M. D. Shermis and J. C. Burstein, *Automated essay scoring: A cross-disciplinary perspective*. Routledge, 2002.
[6] J. Burstein, "Automated essay evaluation and scoring," *The Encyclopedia of Applied Linguistics*, 2013.
[7] C. Ramineni and D. M. Williamson, "Automated essay scoring: Psychometric guidelines and practices," *Assessing Writing*, vol. 18, no. 1, pp. 25–39, 2013.
[8] C.-U. Lei, H.-N. Liang, and K. L. Man, "Advancements in using a machine design project for teaching introductory electrical engineering," in *Teaching, Assessment and Learning for Engineering (TALE), 2013 IEEE International Conference on*, 2013, pp. 556–559.
[9] C.-U. Lei, H. K.-H. So, E. Y. Lam, K. K.-Y. Wong, R. Y.-K. Kwok, and C. K. Chan, "Teaching introductory electrical engineering: a project-based learning experience," in *Proc. IEEE Intl. Conf. on Teaching, Assessment and Learning for Engineering*, 2012, pp. 335–339.
[10] C.-U. Lei, K. L. Man, H.-N. Liang, E. G. Lim, and K. Wan, "Building an intelligent laboratory environment via a cyber-physical system," *International Journal of Distributed Sensor Networks*, vol. 2013, 2013.
[11] C.-U. Lei, K. L. Man, and T. O. Ting, "Using coh-metrix to analyse writing skills of students: A case study in a technological common core curriculum course," in *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2014, IMECS 2014, 12-14 March, 2014, Hong Kong*, pp. 823–825.
[12] A. C. Graesser, D. S. McNamara, and J. M. Kulikowich, "Coh-metrix providing multilevel analyses of text characteristics," *Educational Researcher*, vol. 40, no. 5, pp. 223–234, 2011.
[13] E. Mayfield and C. Rosé, "Lightside: Open source machine learning for text accessible to non-experts," *Invited chapter in the Handbook of Automated Essay Grading*, 2012.
[14] M. D. Shermis and J. Burstein, *Handbook of automated essay evaluation: Current applications and new directions*. Routledge, 2013.
[15] C.-U. Lei, "Exploiting implicit information from data for linear macromodeling," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 3, no. 9, pp. 1570–1577, Sept 2013.