# Automated Essay Scoring with Ontology based on Text Mining and NLTK tools

Jennifer O. Contreras
*Al Madinah International University*
*11th Floor, Plaza Masalam 2, Jalan Tengku Ampuan Zabedah, E&9E 40100 Sha Alam, Selangor, Malaysia*
*jen.almarasy@gmail.com*

Shadi Hilles
*Al Madinah International University*
*11th Floor, Plaza Masalam 2, Jalan Tengku Ampuan Zabedah, E&9E 40100 Sha Alam, Selangor, Malaysia*
*shadihilless@gmail.com*

Zainab Binti Abubakar
*Al Madinah International University*
*11th Floor, Plaza Masalam 2, Jalan Tengku Ampuan Zabedah, E&9E 40100 Sha Alam, Selangor, Malaysia*
*zainab.abubakar@medie.edu.my*

*Abstract*— **One of the common learning activities used in educational levels and disciplines is essay writing. The problems of the essay writing activities are time-consuming, concerns in producing immediate result and/or feedback from teachers to students, and the teachers tend to be subjective in grading the essay activities. The study aims to apply the preliminary approach for automatically generating the domain concept ontology in essays using OntoGen and applied natural language processing algorithms using NLTK (Natural Language Tool Kit) that enhance the teachers essay grading.**

***Keywords—automated essay scoring; natural language processing; ontology; text mining***

## I. INTRODUCTION

Essay type of exam is one of the most essential testing activities at all educational levels. It is an effective tool for assessing academic achievement, integration of ideas and ability to recall [1]. Essays are useful for the evaluation of the learning outcomes of students since it gives student an opportunity to demonstrate their range of skills and knowledge, that includes higher-order thinking. Though essay questions are advantageous to student learning and assessment [2], various challenges for the part of the teachers were noticeable. Manual grading of essays takes up a significant amount of teacher's valuable time [3] since grading essay type exam is time consuming and tedious thing to do especially for a large population of students. Furthermore, the perception of subjectivity of the grading process can be considered since the subjective nature of essay assessment may lead to variation of different results.

Automated Essay Scoring (AES) is a tool that enables teacher to save their time and effort, provide more objective evaluations and refrain from being subjective. Its main goal is to predict the grade of students automatically by means of various features. In the recent years, several essay scoring software were introduced based from different algorithms that includes machine learning which is considered as its core component. Several models, developed approaches, solutions, applications and add-ins have hence been by researchers and software vendors alike for academic and commercial purposes [4].

Text mining is the discovery of interesting knowledge and summarized data in text documents. It is a challenging issue to find accurate knowledge, information (or features) in text documents to help users to find what they want and what data they want [5].

Ontology is a formal explicit specification of the terms in the domain and the relations among them (Gruber, 1993), it is a study about how entities can be grouped and related within a hierarchy [6]. Entities can be subdivided based on distinctive and commonly occurring features. In this study, ontologies can be used to identify the concepts needed for a specific topic in an automated essay scoring.

The proposed approach employs ontology-based information extraction which uses basic natural language processing [7] algorithms for tokenization, word tagging, counting characters, words and sentences, frequency distribution and semantic matching of text for the automatic essay scoring. In this study, we aim to identify if using ontology is useful for the teachers in computing the essay grades of students and explore different natural language tools to extract features for grading the students essay.

## II. DATA

### A. Corpus

Originally, we intend to use the proposed automated essay scoring for the nursing students of Faculty of Nursing in Benghazi University. Unfortunately, we were not able to gather essay corpus yet due to its current situation. For this study, we used real essay exams from the first year Computer Science students at Far Eastern University Institute of Technology. Our corpus is composed of different essays about Human Computer Interaction. Students were asked to answer an essay exam about "Human Factor" topic in Human Computer Interaction course. The essay set has an average essay length of 1238 characters, 118 words and 8 sentences.

## III. METHODOLOGY

We divide the stages of the study into two phases, the first phase includes the use of OntoGen to generate ontologies and determine the similarity from a corpus of relevant essays in each domain. The second phase is extracting features using NLTK tools that can be used for scoring the essay.

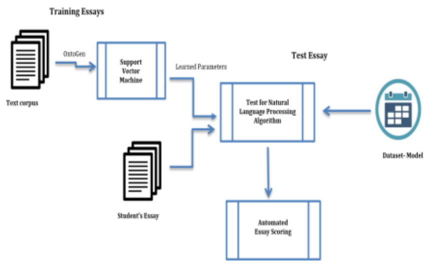


Fig. 1. Implementation of Methodology

Figure 1 shows the implementation of our study, it starts with the extraction of domain ontology from a text corpus. OntoGen will help the teachers to identify the concept domains and the similarity of the essays. Then we used the learning model to determine the parameters based on the features extracted. Scores were predicted for a distinct set of test essays. We chose features that may serve as a substitute for what a human grader might look for in grading students essay.

### A. *Generating Domain Ontology*

Ontology can be developed thru different tools like Protégé', Text2Onto etc., but for our study we used OntoGen. OntoGen is a semi-automatic and data-driven ontology editor focusing on editing of topic ontologies (a set of topics connected with different types of relations) [8]. It is a blend of text-mining techniques and a good user interface to speed up the development of ontology. Thru this, it can bridge the gap between the complexity of using ontology editing tools and the domain experts who are developing the ontology but do not have enough knowledge about ontology engineering.

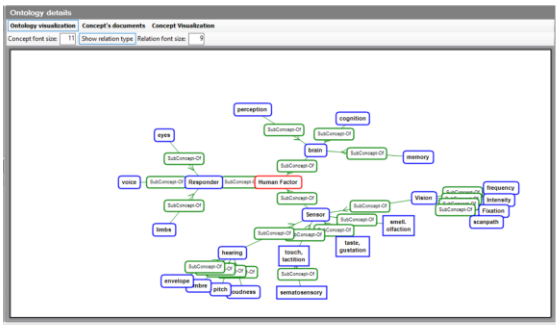

Fig 2. The concepts generated using supervised learngin and ontology visualization

Here, we assumed that the teacher provides a corpus of relevant documents in the domain of interest. The teachers can check from textual information sources of the domain such as textbooks, lecture notes or any.

Figure 2 shows ontology visualization of the domain ontology generated by OntoGen that was also edited by the teacher for enhancement purposes. The teacher used query to tag the essay related to the concept that needs to be added in the ontology.

### B. *Evaluation of Corpus using OntoGen*

OntoGen is using Support Vector Machine (SVM) algorithm to identify the keywords. It suggests concepts based on the list of documents that are currently selected. The concept hierarchy is rooted in a root node, and sub-concepts are treated as children in the ontology tree.

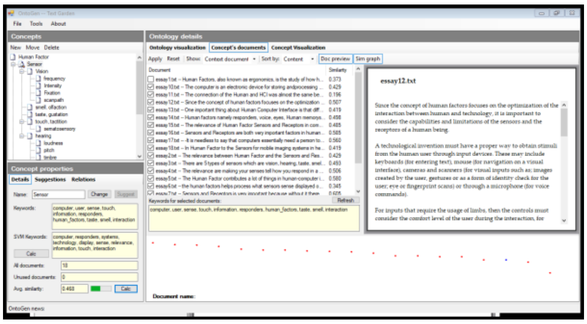

Fig 3. OntoGen Ontology Concept's documents view

Figure 3 illustrates the concept documents based on the trainings done in OntoGen. The concept suggestion was based on clustering into a given number of clusters. The visualization of the documents in the corpus supports the user during the ontology creation. Important clusters with many documents can be recognised easily by the density in the visualization, and this can give hints about how to split the corpus (or already obtained clusters) in the ontology-learning phase.

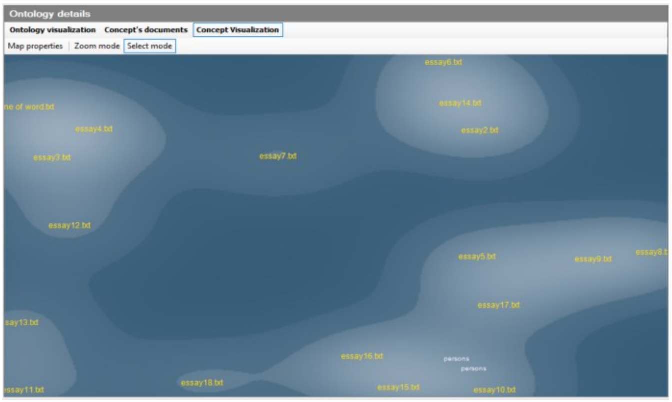

Fig 4. OntoGen Ontology Concept's visualzation view

Figure 4 shows the ontology generated from using supervised learning. This will help the teachers to select which ontology they want to use for scoring their essay.

### *C. Linear Regression*

Linear regression is a linear approach to modeling the relationship between a scalar [wiki] dependent variable Y and one or more explanatory variables (or independent variables) denoted by X.

The X represents the features extracted from the essays and the Y represents the corresponding score for training the model. We have used scikit-learn library in python to implement the linear regression model.

## IV. FEATURES

Feature extraction can be considered as one of the most important part of any machine learning task. In this study, we used OntoGen and NLTK libraries for text processing and feature extraction. Our goal is to build effective essay scoring algorithm by using model attributes like essay length, word counts, correctness, vocabulary and types of word used, domain information etc.

We extracted features based on the following categories to train our model.

### *A. Domain ontology*

We used this as a basis to generate concepts that are relevant to the subject, this is a measure of the presence of specific content and concepts in each essay set that the teacher is looking for. The teacher will use OntoGen's supervise learning method to generate the domain ontology.

### *B. Numerical features*

These are the features like the total word count and sentence county per essay that includes the following algorithms:

- *Tokenization* is the process of splitting or segmenting sentences into their constituent words. A sentence is a collection of words, and with tokenization we essentially split a sentence into a list of words that can be used to reconstruct the sentence. [9].

- *Stop words* is a commonly used word (such as "the", "a", "an", "in") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. In fact, removing stop words is more favorable to have a better result. We would not want these words taking up space in our database, or taking up valuable processing time.

### *C. Parts of Speech count*

Part of speech count is based on classifying words and labeling them based on the different parts of speech such as nouns, adjectives, adverbs and verbs that is helpful to identify student's vocabulary skills. We have extracted the frequency count of different parts of speech in the essay using NLTK-parts-of-speech tagger. Essay were tokenized into sentences before the tagging process.

### *D. Orthography*

Checking for the correct spelling also contributes to a good essay. For this, we extracted number of spelling errors per essay using Pychant library which is spelling correction API used in python.

### *E. Similarity*

Some datasets have a source essay based on which a question has been asked [10]. However, essay type exam is not limited to have one correct answer that is why we have compared the similarity between the source essay and the answered essay. We used Latent Semantic Algorithm (LSA) to compare the similarity of the essays.

LSA is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text [11]. It is a technique in natural language processing of analyzing relationships between documents and the terms they contain by producing a set of concepts related to the documents and terms [10]. It presumes that those words with nearly identical meaning will produce similar pieces of text.

OntoGen uses cosine similarity measure that is commonly used in information retrieval and text mining to determine the semantic closeness of two documents where document features are represented using Bag of Words vector space model, used to position the document according to their similarity to the representative document of a selected domain [12].

Thus, the similarity of both paragraphs can be represented by identifying the high value of cosine similarity. Given two vectors of attributes, A and B, the cosine similarity *cos(θ)* is represented using a dot product and magnitude as:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

Where $A_i$ and $B_i$ are components of vector $A$ and $B$ respectively.

The similarity scales between 0 and 1 as the maximum value which means:

- Cos $0^0 = 1$, $0^0$ means that the two documents are equal, since two sequences point to the same point

- Cos $90^0 = 0$, $90^0$ means that the two documents are totally different.

We can simply say that the bigger the return value is, the more similar the two texts are.

## V. RESULTS AND ANALYSIS

After we finished extracting the features and created the domain ontology, we use linear regression to predict the scores of each essay.

### *A. Supervised learning vs unsupervised learning*

One of the objectives of this study is to determine which is the best method in generating domain ontology.

**Table 1. Average similarities for the essay set**

| *Type of Learning* | *Similarities* |
|---|---|
| Supervised | 47.16% |
| Unsupervised | 34.65% |

Table 1 shows that the average similarity for supervised learning is 47.16% out of the 19 essays as compared to the unsupervised learning which is 34.65%, therefore, we can say that supervised learning is better than unsupervised learning in terms linear regression.

### *B. Similarity result from OntoGen*

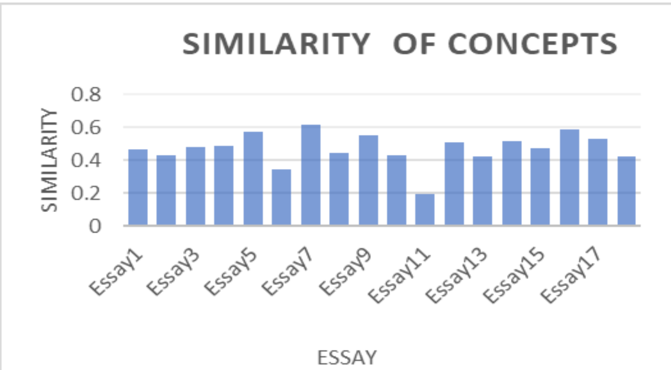


Fig. 5 Similarity of Conceprs extracted from the essay

Figure 5 shows the similarity of the concepts extracted from the essay using OntoGen. It is obvious that in a supervised learning paradigm, the participation of a teacher who provides learner with a piece of training data is better than unsupervised learning where there is no place for a teacher to make prior correct decisions.

### *C. Feature Analysis*

For now, we only used one data set but once we gathered essay exams from the University of Benghazi, we will test more data sets. Afterwards, we have implemented linear regression model using features and documented result.

**Table 2. Average Character, Word and Sentence Count**

| *Essay* | *Character Count* | *Word Count* | *Sentence Count* |
|---|---|---|---|
| Essay1 | 1995 | 194 | 11 |
| Essay2 | 128 | 11 | 1 |
| Essay3 | 1096 | 112 | 18 |
| Essay4 | 1208 | 123 | 16 |
| Essay5 | 1413 | 140 | 5 |
| Essay6 | 160 | 15 | 1 |
| Essay7 | 1295 | 117 | 12 |
| Essay8 | 153 | 11 | 1 |
| Essay9 | 228 | 21 | 2 |
| Essay10 | 1103 | 97 | 6 |
| Essay11 | 1202 | 139 | 8 |
| Essay12 | 3279 | 299 | 16 |
| Essay13 | 1338 | 128 | 7 |
| Essay14 | 1585 | 150 | 4 |
| Essay15 | 2221 | 188 | 8 |
| Essay16 | 945 | 99 | 7 |
| Essay17 | 1775 | 167 | 12 |
| Essay18 | 1157 | 112 | 6 |
| Average | 1238 | 118 | 8 |

In this study, we hypothesized that counting of words from the essay would certainly be correlated positively with the essay score. We extracted the total character, words and sentences as represented on the table 2, however we only considered "word count" against the human score since words correlates to the vocabulary.

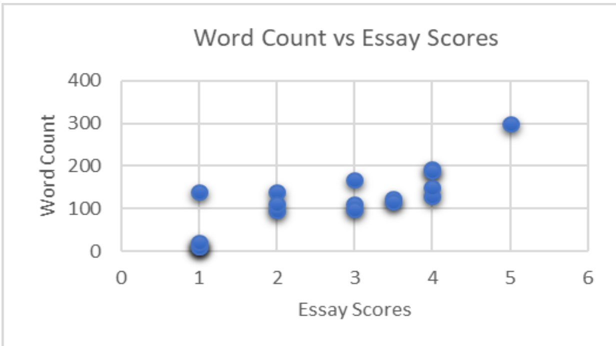


Fig. 6 Word Count vs Essay Score dataset

Figure 5 is a scatterplot that shows a strong, positive and linear association between word count and the essay score dataset. The essay scoring that was based on the TOEFL writing rubric where 6 is the lowest and 0 is the highest. The figure implies that a bigger number of word count results, can contribute a higher probability of getting a high essay score.

We can say that a student who are more exposed to a larger vocabulary list and comprehend its uses can write better essay. Furthermore, the vocabularies can also be correlated to the domain ontology thru its concept. Therefore, we hypothesized that a larger vocabulary list would correlate with a higher essay score [13].

To strengthen our claim, we also count the frequency distribution of the words used in an essay who got the highest score and the lowest score.

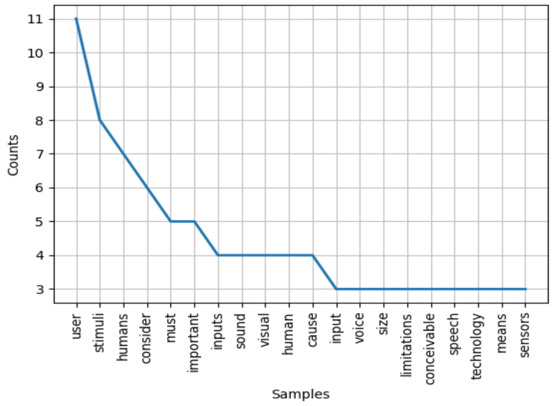


Fig. 7. Cumulative Frequency Plot for 20 most frequently words in Essay 12: after deleting stopwords and punctuations

Figure 7 shows the cumulative frequency distribution for 20 most frequently words extracted from essay12 which obtains the highest score, therefore, we can say that the bigger word counts the higher the score you can get.

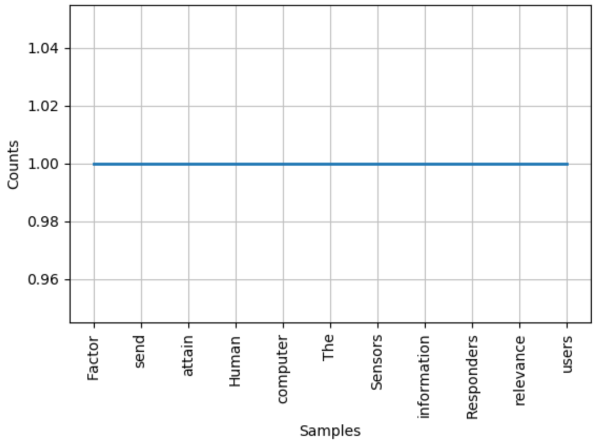


Fig 8. Cumulative Frequency Plot for 20 Most Frequently Words from the lowest score Essay 3: after deleting stopwords and punctuations

Figure 8 shows the cumulative frequency plot for 20 most frequently words from the lowest score. It is obvious that the students have few words used therefore the score is low.

Counting the misspelled words is also one of the features we extracted for scoring the essay.

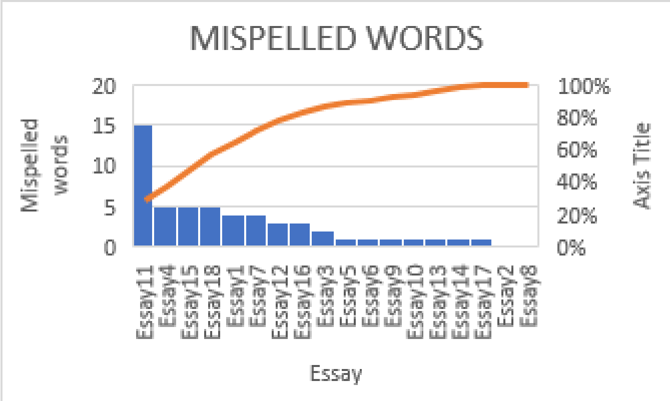


Fig 10. Counting of misspelled words from the essay

Figure 10 shows the summary of the counting of the misspelled words from the essay.

## VI. CONCLUSION AND FUTURE SCOPE

We proposed an automatic essay grading with machine learning application using OntoGen for generating the domain ontology which we found useful especially to the teachers who doesn't know how to use different tools in creating their own ontology. The results obtained seem quite encouraging especially the automatic computation of the similarity among documents. We also proved our hypothesis that the number of words or vocabulary the students can exhibit contribute to the score of the essay.

For further studies, we will add recommender to help students enhance their writing skills, enhanced the rubric in computing the concept extracted from the ontology and add sentiment analysis.

## REFERENCES


[1] S. Drolia, P. Agarwal, S. Rupani and A. Singh, "Automated Essay Rater using Natural Language Processing," International Journal of Computer Applications (0975-8887). Vol. 163 – No 10, April 2017.

[2] G. Adrian, B. and S. Roxana, "Enacting textual entailment and ontologies for automated essay scoring in chemical domain," 16th Int. Symposium on Computational Intelligence and Informatics (CINTI2015), Budapest, Hungary, arXiv:1511.02669v1, 19-21 November, 2015.

[3] M. Manvi, J. Mishels, and A. Aswin, "Automated Essay Grading Using Machine Learning". Standford University, December 2012.

[4] M. Fortuna, D. Grobelnik, and D. Mladenic, "OntoGen: Semi-automatic Ontology Editor," HCI International 2007. Beijing, July 2007.

[5] A. Abejide, W. Ibiba, and A. Juliet Chioma, "An Expert System for Automated Essay Scoring (AES) in Computing using Shallow NLP Techniques for Inferencing," International Journal of Computer Applications(0075-8887). London, vol. 51 – No. 10, August 2012.

[6] G. Chakraborty, M. Pagolu and S. Garla, "Text Mining and Analysis Practical Methods, Examples and Cases Studies Using SAS," HCI International 2007. Beijing, July 2007.

[7] O. Daramola, I. Afolabi I. Akinyemi and O. Oladipupo, "Using Ontology-based Information Extraction for Subject-based Auto-grading," IC3K 2013, KEOD 2013- 5th International Conference on Knowledge Engineering and Ontology Development. Proceedings. 373-378. 2013.

[8] Fortuna, B., Grobelnik, M., & Mladenic, D. (n.d.). OntoGen Semi-automatic Ontology Editor. Retrieved June 2, 2018, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.9381&rep=rep1&type=pdf

[9] D. Saknar, "Text Analytics with Python," Apress Publishing, 2016. p.112 [online] Available: Google books

[10] S. Drolia, P. Agarwal, S. Rupani and A. Singh, "Automated Essay Rater using Natural Language Processing," International Journal of Computer Applications (0975-8887). Vol. 163 – No 10, April 2017.

[11] T. Landauer, P. Foltz, and D. Laham,. " Introduction to Latent Semantic Analysis. Discourse Processes", **25**, 259-284, 1998.

[12] M. Berthold , "Bisociative Knowledge Discover: An Introduction to Concept, Algorithms, Tools, and Applications", p. 318 June 2012

[13] Gupta, A., Hwang, A., Lisker, P., & Loughlin, K. (2016, December 14). Automated Essay Grading. Retrieved June 02, 2018, from https://kevinloughlin.github.io/Automated-Essay-Grading/