Descriptive Answer Evaluation using NLP Processes Integrated with Strategically Constructed Semantic Skill Ontologies

Gerard Deepak[1],*, Ayush Kumar A[2], Sheeba Priyadarshini[3] and Divyanshu Singh[4]

[1]*Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India*
[2]*National Institute of Technology, Tiruchirappalli*
[3]*CHRIST(Deemed to be University), Bangalore;*
[4]*Birla Institute of Technology and Science, Pilani*

Abstract

The world is moving towards an online methodology of education. One of the key challenges is the assessment of questions which do not have a definite answer and have several correct answers. To solve this problem, and for quality evaluation of descriptive answers online, an automatic evaluation methodology is proposed in this work. A language model is modelled from the expected answer key, and entity graphs are generated from the ontology modelled using the input answer to be evaluated. Natural Language Processing (NLP) techniques like Stemming, Summarization, and Polarity Analysis are integrated in this work with Ontologies for the efficient evaluation of descriptive answers. Several challenges which come across evaluating descriptive answers are discussed in this chapter, and they have been solved in order to obtain a dynamic and robust evaluating system. Finally, the system is evaluated using a user-feedback methodology comprising a panel of 100 students and 100 professors.

Keywords

E-Learning, Keyword Extraction, Natural Language Processing, Online Evaluation, Ontology

1. Introduction

Many students take online exams and classes for certifications in this highly competitive world. There are many challenges associated with the online mode of education. Some of them are the availability of an internet connection, doubt clearing methods, real time classes, and evaluation of assignments. To solve the problem of evaluation, most of the assessments online are in Multiple Choice Questions (MCQ) form, which involves one or more correct answers for each question. The limitation with this form of assessment is the definite nature of the questions, which do not kindle creativity in the learners and are based on a rigid system of questions and answers. The usage of MCQ-based evaluation became inevitable due to the complicated nature of descriptive-type answers, and their evaluation. This chapter proposes

IWMSW-2022: International Workshop on Multilingual Semantic Web, Co-located with the KGSWC-2022, November 21–23, 2022, Madrid, Spain

*Corresponding author.

gerard.deepak.christuni@gmail.com (G. Deepak); ayush.kumar376@gmail.com (A. K. A); pdj.sheeba@gmail.com (S. Priyadarshini); divyanshu9676@gmail.com (D. Singh)
0000-0003-0466-2143 (G. Deepak)

a robust solution to this problem, as NLP-based techniques are integrated with Ontologies to program a Descriptive Answer Evaluator System (DAES). This DAES system used ontologies in place of language models and hence makes the entire process efficient. Entity graphs are developed using ontologies, and it is compared with the language model for further evaluation. NLP based techniques are used to compare the answers with the answer key, and keywords are generated by summarizing the answer key.

1.1. Motivation

Several online platforms offer online courses and evaluation, but most of the exams and assessments are either Multiple Choice Questions (MCQ) based because they are simpler to evaluate and require less complexity to assess. This hinders the learning process and students are not able to implement their knowledge gained from the course effectively as many courses are incomplete without the questions of the types “What-if”, “Compare and Contrast”, “Describe”, etc. The problem in assessment of these questions is the requirement of a human intermediary for evaluation of the answers. This chapter aims to provide a solution to the problem by modeling an efficient and robust descriptive answer evaluation system.

1.2. Contribution

Online evaluation of descriptive answers has many challenges like tacking spelling mistakes, evaluating semi-complete answers, not able to decode the mindset of the examinee, and many others. Hence, to decode all these and to make an efficient evaluator, a language model should be used which must be relevant to the subject matter. An ontology is used for properly structuring the components of the extracted text, and entity graphs are used for comparing. Finally, several NLP techniques like Semantic similarity analysis, Polarity analysis are used for effective paper grading. The marks are allotted based on similarity of words, and exact word match, as an answer with proper use of jargons is considered better than those with non-technical words.

1.3. Organization

The remainder of the chapter is organized as follows: Section 2 consists of the relevant Literature Review. In Section 3, the System Architecture is described. Ontology modeling and conceptualization are dealt with in Section 4. Section 5 consists of the Implementation, Results, and Discussions. The chapter is finally concluded in Section 6.

2. Literature Review

Pawade et al., [1] have proposed a model for question answering which is an Open Domain model and incorporates ALBERT pretrained models with variational parameter sizes which takes care of content-context mapping. This model is open domain and targets question answering as an Open Domain problem and provides an NLP oriented solution. Nandini et al., [2] have proposed a DAES which comprises of several stages like question and answer classification and grading. The problem of dealing with neutral language answers has been solved by precise meaning extraction which leads to appropriate grading. Also, a cognitive-based approach is adopted in this system. Kapoor et al. [3] have presented a systematic study on the present DAES like n-gram models and other parameters which are crucial for the evaluation of answers automatically. Dubey et al. [4] have proposed a DAES which evaluated 96.22% of answers with a set threshold. This classifier works on the principle of random forest and works on 530 training samples of evaluating descriptive answers automatically. Hussain et al. [5] have presented a digital evaluating system for Bangla scripts using keyword generation and a search on the generated keywords. Vinothina et al. [6] have proposed an EVaClassifier for automatic evaluation of descriptive answers using Support Vector Machines (SVMs). The evaluation of the proposed system is done by the accuracy of grading by the supervised machine learning algorithm proposed. Meena et al., [7] have devised a method for DAES using Hyperspace Analog to Language techniques building a Self-organizing Map. Then the input is clustered and the accuracy of results increases significantly for the evaluation. Kudi et al., [8] have proposed an evaluation system focussing on short text matching using NLP techniques like text mining, knowledge distillation etc. Using QAML (Question Answer Markup Lanugage), they have defined a structure for the answers and compared between JSON and XML. Kuzi et al. [9] have proposed a system for automatic assessment of descriptive answers using deep features like polysemy and synonymy. They have evaluated their method on a dataset by vetinery students and have concluded that the quality of results can be incrwased by pooling several techniques together and forming a larger feature set. Usip et al. [10] have also putforth a personal profile Ontology to ease Software Requirements Engineering allocation of tasks. The proposed Ontology model captures both static and dynamic data properties and also mixes Ontological Strategies like Neon and Methontology along with e-PPO model for achieving dynamism over ontological properties for task allocation and reasoning. Rai et al., [11] have put forth strategies for conceptual modeling of Ontologies where a novel conceptual model has been put forth which focuses on the management aspects along with representational aspects for assimilating domain knowledge where entities, relationships along with attributes are used to correlate domain centered knowledge. Tiwari et al., [12] have proposed a semantic modeling paradigm focusing on healthcare as a core domain of choice. They have formalized ontology as a backbone of Semantic Web and built semantic models with a formal explicit specification for healthcare along with connected IoT devices. The framework focuses on semantic annotation, linking and modelic along with representational aspects of Ontologies and connected IoT devices. Patience et al., [13] have put forth personal profile ontology focusing on personnel appraisals to justify the fact that encompassment of smart resume reduces the risk of having a large number of tasks which consumes more processing time units. The model uses a mix of methontology and Neon paradigms alsong with the e-PPO strategy for enabling constraint driven evaluation over ontologies to overcome personal bias. Panchal et al. [14] have formulated a semantic model which encompass Ontologies for a higher education system offered by public universities. The university ontology AISHE-Onto is explicitly formalized where SPARQL querying have been applied for realization of reasoning on the Ontology put forth. ECODO Ontology put forth by Fernando et al. [15] is an example for representing and sharing knowledge for legal documentation in e-governance applications over the Web 3.0. These literatures indicate that the domain Ontology representation and authoring can lead to a reasonable semantic inferencing for ensuring Web 3.0 compliant applications exhibit strong degree of collective intelligence and

Ontologies ensure inferential capabilities for deriving reasonable and sharable knowledge.

3. Proposed System Architecture

The system architecture of the DAES system is shown in Figure 1. The DAES system consists of many steps for the efficient computation of the score of the answer as compared to the answer key. The robust tool must be able to avoid the True Negatives, as there should not be any correct word deemed as incorrect. The Text is extracted from the resume database, and it is taken through a preprocessing step, to make the text efficient and ready for ontology modeling. The text is detected from the PDF document, which was uploaded, and it is segmented into several sentences. The individual words are tokenized, and the stop words are removed. Stemming of the words happen, and then the integrated set of words are used to automatically model the ontology. After the generation of ontology, entity graphs are formed, and they are compared with the Language model which is developed using the Answer Database by summarizing the content and arranging the data hierarchically. Semantic comparison of the keywords then takes place, as the match based on identical matches are given a higher weightage, as they indicate the learning of the student during classes. Then the count of matching words is determined using TF-IDF calculation and several Similarity index computations. The Polarity analysis with respect to the answer is done, and the final score is calculated. The feedback along with the scores is displayed in the output.

3.1. Text Preprocessing

Similarity between the sentences can be calculated using cosine similarity. The terms σ in the sentences are weighted using Tf-Isf shown in Equation (1) and Equation (2) respectively. The term frequency in the given document is the number of times a given term appears in the document:

$$Tf_x = \frac{T_x}{\Sigma_{x=1}^{len} T_n} \tag{1}$$

where T_x denotes the frequency of the word and T_n denotes the sum of occurrences of all the word in the document. The measure of the importance of the term is calculated by a factor called inverse sentence frequency using the Equation defined in (2).

$$ISF = log(\frac{C}{\lambda_n}) \tag{2}$$

where C denotes the total count of sentences in the document and is the number of sentences containing the significant term. The corresponding weight is, therefore, computed as in Equation (3).

$$We_\sigma = Tf_{nm} * ISF_n \tag{3}$$

where Tf_{ij} is the term frequency of nth index term in mth sentence and ISF i is of inverse sentence frequency. The similarity of the sentence can be measured using cosine similarity as in Equation (4).

$$cos(\sigma_n, \sigma_m) = \frac{\sigma_n . \sigma_m}{||\sigma_m|.|\sigma_n||} \tag{4}$$

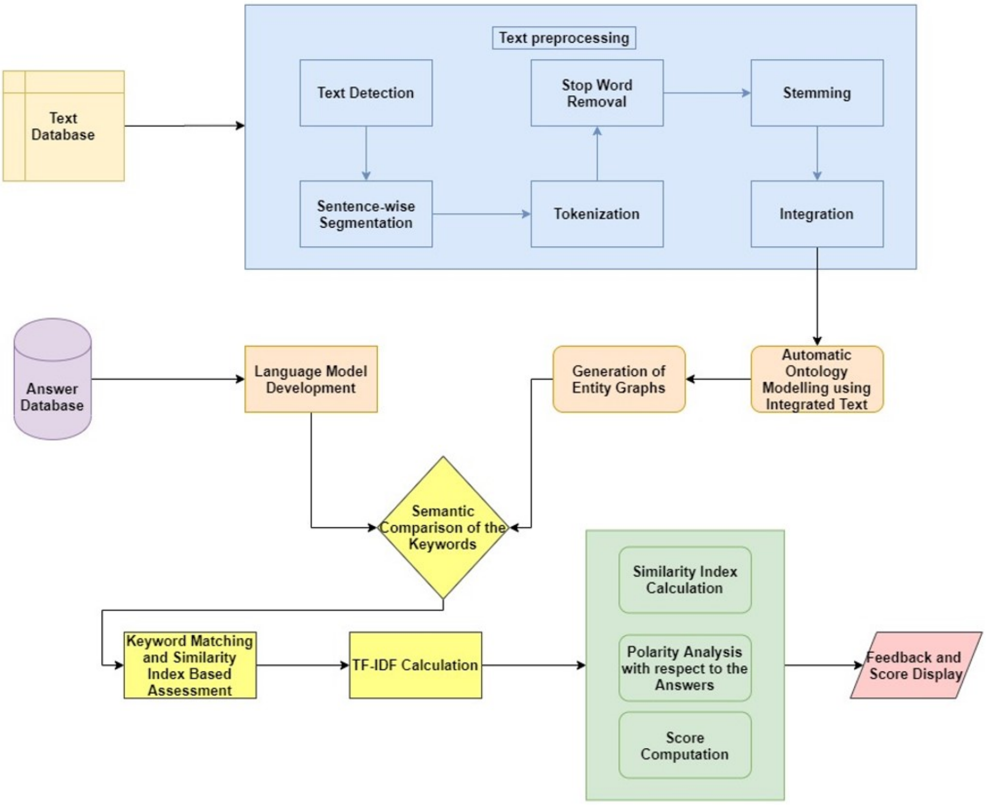
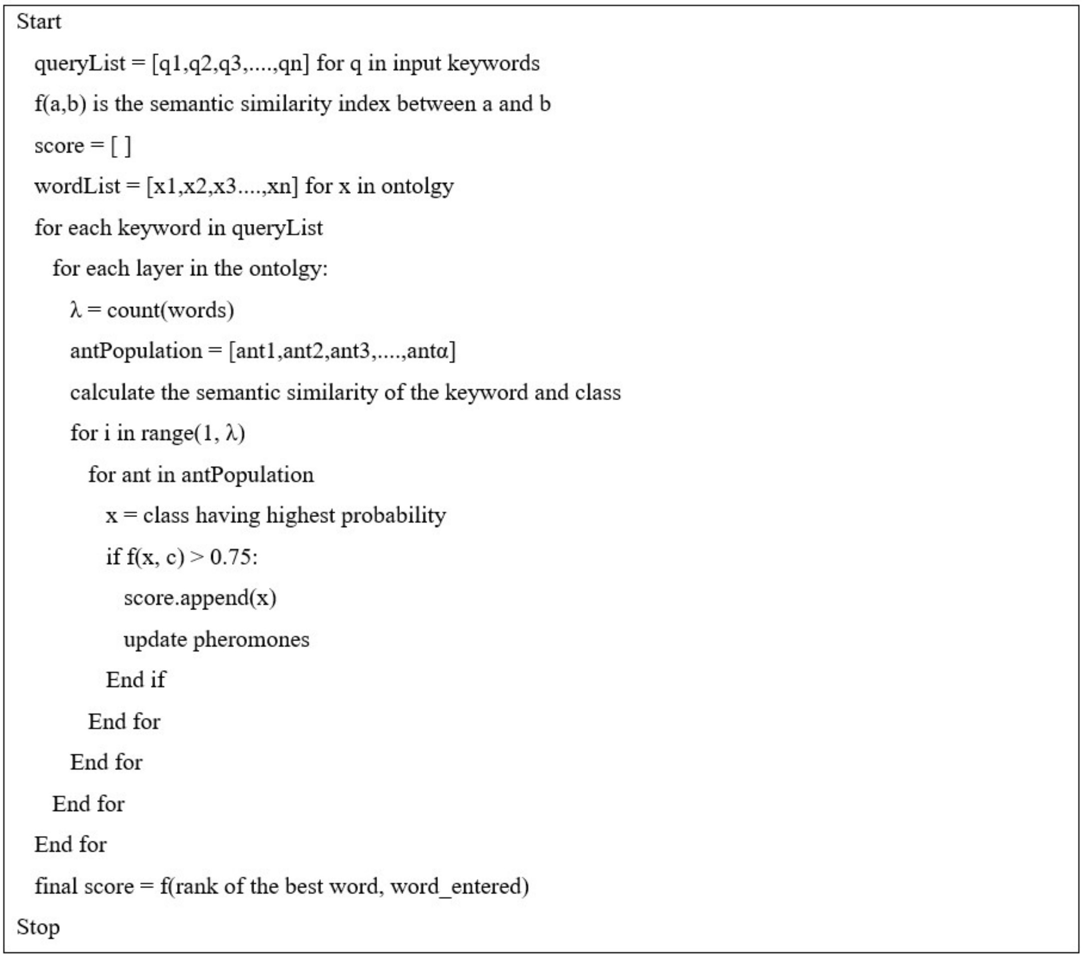

Text Database
Text preprocessing
Text Detection
Sentence-wise Segmentation
Tokenization
Stop Word Removal
Stemming
Integration
Answer Database
Language Model Development
Generation of Entity Graphs
Automatic Ontology Modelling using Integrated Text
Semantic Comparison of the Keywords
Keyword Matching and Similarity Index Based Assessment
TF-IDF Calculation
Similarity Index Calculation
Polarity Analysis with respect to the Answers
Score Computation
Feedback and Score Display

Figure 1: System Architecture

3.2. Ontology Modeling

In today' world. many open-access evaluation tools are available for use which evaluate the responses of the examinees. But they all have limited reach, as they can evaluate only MCQs, Numeric answer type questions, or short one-line type questions. One of the limitations of such tools is that the answer order gets lost during summarization and comparison. So, for a question which requires step-by-step procedure, even a wrong answer is given full marks. This kindles the need for the integration of hierarchy in the language models used to compare the Student response and the correct answer. Such tools underassess the answer, and the essence of E- learning is lost. Such low-level assessment can also result in faulty grading of highly competent answers. The word demand, and the vocabulary level is not assessed in such tools. The usage of ontology in this case gives a hierarchical ordering, and asserts perfect grading in every case, as a strategic tool is very much required for grading. From the preprocessed text, an ontology is generated using the keywords, and a hierarchically structured data form is obtained. A strategic ontology with attributes as the synonyms and the part of speech of the word is used for this work.

Figure 2: Algorithm for DAES



```
Start
  queryList = [q1,q2,q3,....,qn] for q in input keywords
  f(a,b) is the semantic similarity index between a and b
  score = [ ]
  wordList = [x1,x2,x3....,xn] for x in ontolgy
  for each keyword in queryList
    for each layer in the ontolgy:
      λ = count(words)
      antPopulation = [ant1,ant2,ant3,....,antα]
      calculate the semantic similarity of the keyword and class
      for i in range(1, λ)
        for ant in antPopulation
          x = class having highest probability
          if f(x, c) > 0.75:
            score.append(x)
            update pheromones
          End if
        End for
      End for
    End for
  End for
  final score = f(rank of the best word, word_entered)
Stop
```

4. Implementation

The algorithm for DAES has been designed using the Ant Colony Optimization and depicted as Figure 2 and implemented in Windows 10 operating system. The implementation was done on a computer with Intel i5 processor and includes 8GB of RAM. Pre-processing is the first step in the implementation of the algorithm. The text is detected from the PDF using a suitable tool. Then, sentence wise text segmentation is done. Segmentation is the splitting of the entire answers into individual sentences, which can be then easily processed. Doing this also ensures that no information is missed out during summarization. The words are tokenized, and all the unnecessary words, which render no useful information to the question, and ae present only for grammatical purposes are removed, and the keywords are extracted. Stemming of words is followed by integration of all the data and making it ready for the next step which involves Automatic Ontology Modelling.

Ant Colony Optimization is an algorithm that takes insights from the biological phenomena of ants finding their paths from home to their food, and their ability to return back home. This can be easily used for finding the global maxima or the minima of any function. The ants maximize their probability of reaching back home by the emission of pheromone, a chemical compound, which is maximized by all other ants, as they tend to follow the same path, and emit pheromone too. This can be used to traverse an ontology to find the most similar word to the keyword in the student's answer ontology and the language model of the answer key. The semantic similarity function can be taken as the target function and can be maximized by the regular emission of pheromones. The pheromones can be related to the semantic similarity index, and the function is maximized by high emission of this pheromone, or high similarity index in other words. This abstract optimization algorithm can help to calculate the score of the student from the ontology, and hence, fulfils the purpose of the DAES.

The scoring algorithm uses the language model and the ontology obtained from the answer key to assess the answer sheet provided by the student. TF-IDF calculations are performed and a variety of tests are done to ensure proper allocation of scores to the answer sheet. The first form of assessment is the Keyword matching and similarity index-based scoring which carries the maximum weightage as it indicates the presence of perfect or matching words in the students' answer as compared to the answer key generated by the professor. Other types of checks like the polarity tests can also play a dominating role in questions where one has to put forth one's iDAES which can be either in support of or opposing sensitive topics. Hence, for the evaluation of answers which depend on arts, and humanities like subjects, polarity scores must also be computed. The final score is based on the weighted mean of the several types of the metrics used and hence, a perfect score is allocated to the student.

4.1. Results and Evaluation

The output of the DAES is the overall score which is basically dependant on classification of the words, and their detection. Since the core of the DAES is classification of the words into the several probable classes, a confusion matrix can be used to evaluate the results of the system. there may be some words which can be misclassified as not suitable for the particular instance, and some words may be wrongly identified correct even if they are not suitable for the occurrence. The metrics that can be considered for evaluation are Precision, Recall, Accuracy, and the False Positive Rate. Precision is defined as a proportion of the rightly classified words that are truly suitable. Precision is measured by the formula given in Equation (5).

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Recall can be defined as the proportion of the selected words based on the classification w.r.t total number of words considered. is measured by the formula given in Equation (6).

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Accuracy can be described as in Equation (7).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Table 1
*

Performance Measures of the proposed DAES

Number of Instances	Precision%	Recall%	Accuracy%	FDR%
20	95.41	97.01	96.21	96.20
40	94.87	96.43	95.65	95.64
60	93.15	95.45	94.30	94.29
80	92.76	94.91	93.84	93.82
1000	92.02	94.01	93.02	93.00
Average	93.64	95.56	94.60	94.59

The False Positive Rate (FPR) is the number of irrelevant words that are identified as correct by the system (all FPs), divided by the total number of irrelevant words. False Positive Rate (FPR) is defined as follows in Equation (8).

$$FPR = \frac{FP}{TN + FP} \tag{8}$$

Table 1 depicts the precision, recall, accuracy, and the F-Measure of the proposed DescEval algorithm using ACO. The average precision is calculated to be 93.64%, the average recall is 95.56%. For this algorithm, the average accuracy is 94.60% and the average F-Measure of the assessment is 94.59%. These parameters are found to be higher than the state-of-the-art systems which do not use ontologies and rather rely on conventional NLP techniques.

The proposed DescEval system using Ant Colony Optimization Algorithm for semantic ranking and similarity matching with overall grading using semantically aware ontologies performs well on the curated dataset. It has an overall average accuracy of 94.60%. This approach semantically structures the contents of the expected answer key into a dynamic ontology, and then the information is retrieved using ant colony optimization for assessing the answer script used by the student. More focus is given on the similarity between words and both taxonomic and non-taxonomic similarity are considered as in an open book examination, the student must be free to express his thoughts and there must not be any kind of barriers to the language used. All the attributes and intricacies of the dataset (expected answer key) is stored in the form of an ontology because the attributes are more dynamically linked with each other and there is an efficient retrieval system. Figure 3 depicts the variation of precision, recall, accuracy, and F-Measure with the number of answer scripts evaluated for different domains.

Another form of evaluation based on the overall performance of the system is by comparing the grading done by the system to the grading done by experienced professors and evaluators for the same set of answer papers. Five different groups of question chapters are selected, and they are evaluated by five professors for ten different students. The results are shown in the Table 2.

From the results, it is certain that the technique proposed here gives sufficiently low error metrics and it could be used in real life applications. The low variance is due to proper semantic mapping of the keywords and efficient computation of the score which does not judge on only

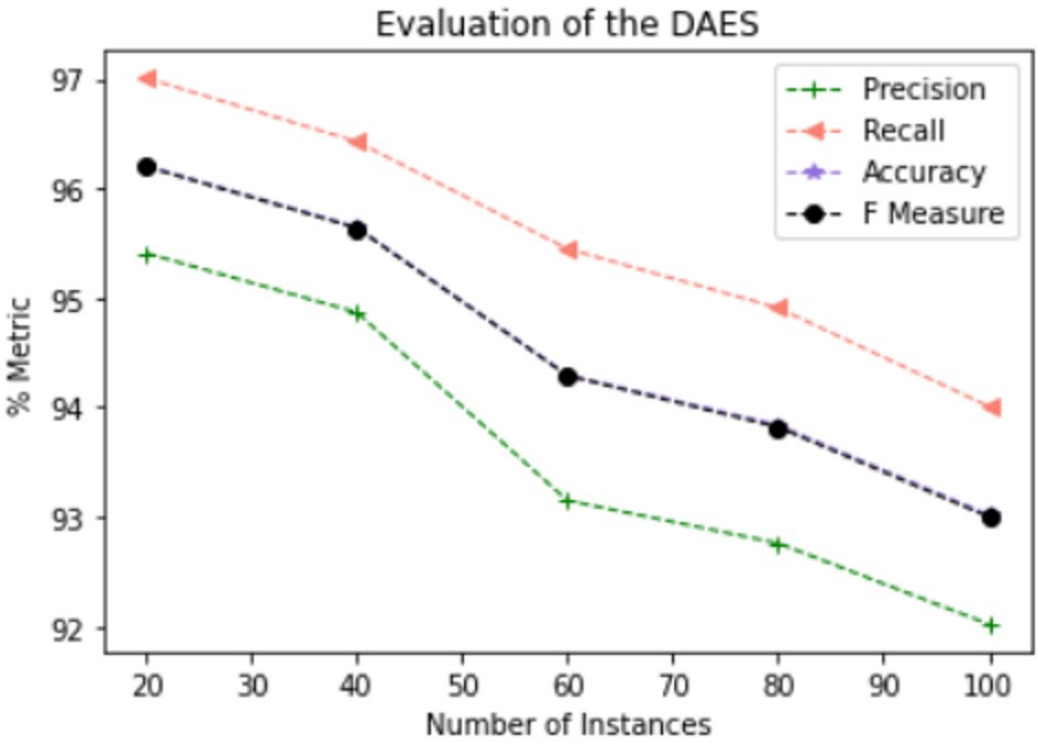


Figure 3: Evaluation of the DAES

Table 2
Performance Comparison of DAES for different sets of question papers

SI. No.	Question paper	Average score given by the Professor	Average Score predicted by DAES	Mean Square Error	Mean Absolute Percentage Error	Mean Poisson Deviance
1	Set 1 - Physics	67.8	69.5	2.80	3.21	0.11
2	Set 2 - Chemistry	72.5	71.3	1.83	2.04	0.04
3	Set 3 - Ethics	83.0	79.0	3.91	4.69	0.18
4	Set 4 - English Literature	72.5	75.4	3.37	4.15	0.15
5	Set 5 - History	67.5	69.0	2.51	3.27	0.09

a single parameter, but also caters to a wide variety of subjects by using different metrics and methods to ensure that the answers are perfectly graded. Figure 3 depicts the % Mertric Vs No. of Recommendations Comparison for the proposed DAES.

Figures 4(a),4(b),4(c),4(d) depicts the Precision vs Number of Instances, Recall Vs Number of Instances, Accuracy Vs Number of Instances and F-Measure vs Number of Instances respectively

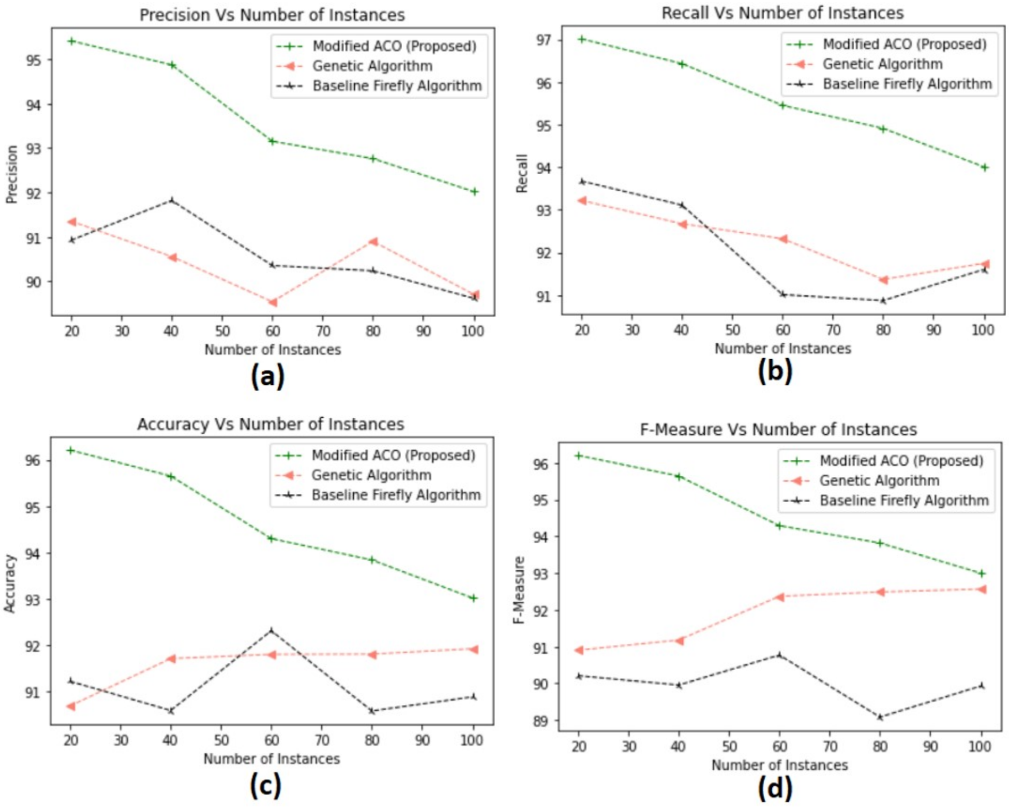


Figure 4: Comparison of Optimization algorithm of DAES with other similar algorithms

for the comparison of the optimization algorithm in the proposed DAES with other similar optimization algorithms. The system was further evaluated by using different optimization algorithms for ontology retrieval rather than the Modified Ant Colony Optimization (MACO) approach used in this chapter. The results show that the average precision is 3.2% higher for the MACO approach when compared to a genetic algorithm-based model, and it is 3.8% higher than that of a baseline firefly optimized model. Similarly, the % increase values for recall are 3.4% and 4.2% respectively. The accuracy is hence 3.3% higher than the genetic algorithm optimized model, and 4.0% higher than the firefly-based model. Also, the F-Measure of MACO approach is found to be 3.3% higher than the genetic algorithm-based model and 3.9% higher than the Firefly optimized paper.

5. Conclusions

The proposed DAES system automates the tedious evaluating system which consumes a lot of time of the evaluators. It is a major challenge for the creators of the online course to assess the online descriptive answer questions in a stipulated period to enable the students to finish the course quickly. The proposed system employs ontologies for structuring of the data and its usage as a language model is perfect as hierarchical structuring improves the accuracy and decreases the chances of underassessment or overassessment of a question. An accuracy of 97.16% is achieved using this DAES. The DAES has been evaluated using the proposed metrics. The results obtained are satisfactory, and hence prove this DAES as best-in-class.

References

[1] Pawade, Dipti, Avani Sakhapara, Isha Joglekar, and Deepanshu Vangani. "Implementation of Open Domain Question Answering System." In International Conference on Data Management, Analytics Innovation, pp. 499-507. Springer, Singapore, 2023.

[2] Nandini, V., & Maheswari, P. U. (2020). Automatic assessment of descriptive answers in online examination system using semantic relational features. The Journal of Supercomputing, 76(6), 4430- 4448.

[3] Kapoor, B. S. J., Nagpure, S. M., Kolhatkar, S. S., Chanore, P. G., Vishwakarma, M. M., & Kokate, R. B. (2020, February). An Analysis of Automated Answer Evaluation Systems based on Machine Learning. In 2020 International Conference on Inventive Computation Technologies (ICICT) (pp. 439- 443). IEEE.

[4] Dubey, R., & Makwana, R. R. S. (2019). Computer-Assisted Valuation of Descriptive Answers Using Weka with RandomForest Classification. In Proceeding of the Second International Conference on Microelectronics, Computing & Communication Systems (MCCS 2017) (pp. 359-366). Springer, Singapore.

[5] Hussain, M. G., Kabir, S., Al Mahmud, T., Khatun, A., Islam, M. J. (2019, September). Assessment of Bangla Descriptive Answer Script Digitally. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-4). IEEE.

[6] Vinothina, V., & Prathap, G. (2019, June). EVaClassifier Using Linear SVM Machine Learning Algorithm. In International Conference on Intelligent Computing and Communication (pp. 503-509). Springer, Singapore.

[7] Meena, K., & Raj, L. (2014, December). Evaluation of the Descriptive type answers using Hyperspace Analog to Language and Self-organizing Map. In 2014 IEEE International Conference on Computational Intelligence and Computing Research (pp. 1-5). IEEE.

[8] Kudi, P., Manekar, A., Daware, K., & Dhatrak, T. (2014, December). Online Examination with short text matching. In 2014 IEEE Global Conference on Wireless Computing & Networking (GCWCN) (pp. 56-60). IEEE.

[9] Kuzi, S., Cope, W., Ferguson, D., Geigle, C., & Zhai, C. (2019, June). Automatic assessment of complex assignments using topic models. In Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale (pp. 1-10).

[10] Usip, P.U., Udo, E.N., Umoeka, I.J. (2021). An Enhanced Personal Profile Ontology for Software Requirements Engineering Tasks Allocation. In: Villazón-Terrazas, B., Ortiz-Rodríguez, F., Tiwari, S., Goyal, A., Jabbar, M. (eds) Knowledge Graphs and Semantic Web. KGSWC 2021. Communications in Computer and Information Science, vol 1459. Springer,

[11] Rai, C., Sivastava, A., Tiwari, S., Abhishek, K. (2021). Towards a Conceptual Modelling of Ontologies. Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, Volume 1, 1286, 39.

[12] Tiwari, Sanju, Fernando Ortiz-Rodriguez, and M. A. Jabbar. "Semantic modeling for healthcare applications: an introduction." Semantic Models in IoT and Ehealth Applications (2022): 1-17.

[13] Usip, Patience Usoro, Edward N. Udo, and Ini J. Umoeka. "Applied personal profile ontology for personnel appraisals." International Journal of Web Information Systems 18, no. 5-6 (2022): 487-500.

[14] Panchal, Ronak, Priya Swaminarayan, Sanju Tiwari, and Fernando Ortiz-Rodriguez. "AISHE-Onto: a semantic model for public higher education universities." In DG. O2021: The 22nd Annual International Conference on Digital Government Research, pp. 545-547. 2021.

[15] Ortiz-Rodriguez, F., Medina-Quintero, J. M., Tiwari, S., Villanueva, V. (2022). EGODO ontology: sharing, retrieving, and exchanging legal documentation across e-government. In Futuristic Trends for Sustainable Development and Sustainable Ecosystems (pp. 261-276). IGI Global.