Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 46 (2025), No. 2, pp. 281–289
DOI : 10.47974/JIOS-1913

Semantic-based enhanced logistic regression model with dual mode classification for accurate MCQs generation

Vijaya Raju Madri *
Sreenivasulu Meruva †
Department of Computer Science and Engineering
Kandula Srinivasa Reddy Memorial College of Engineering
Affiliated to Jawaharlal Nehru Technological University
Ananthapuramu 516003
Andhra Pradesh
India

Abstract

The developments in language models and architectures, such as the transformer, allow us to investigate how teachers benefit from these endeavors. In this research, neural language models are employed to generate Multiple Choice Questions (MCQs) automatically from Wikipedia. MCQs are a common form of assessment. Despite the advantages, creating MCQs manually is difficult, time-consuming, and prone to mistakes. This is because the development of multiple-choice questions requires knowledge from the MCQ developers, as each MCQ consists of a question, a correct answer, and other possibilities. This research presents a method that combines semantic and machine learning techniques to create technically accurate MCQ stems of varying sorts. This paper proposes a Semantic-based Enhanced Logistic Regression Model with Dual Mode Classification (SbELR-DMC) for accurate MCQs and Odd one-out model questions for assessment of students. The proposed model when contrasted with the traditional models exhibits better accuracy.

Subject Classification: *Primary 93A30, Secondary 49K15.*

Keywords: *Multiple choice questions, Machine learning, Semantics, Enhanced logistic regression.*

This research was conducted during the NREUP at Michigan State University and it was sponsored by NSF grants DMS-1156582 and DMS-1359016.

* *E-mail:* **vijayaraju2122@gmail.com (Corresponding Author)**
† *E-mail:* **meruva7@gmail.com**

1. Introduction

It is assessment processes that automatically gauge students' progress throughout the process of learning and instruction are necessary for new e-learning approaches. Tailored can comprehend and raise the caliber of their educational experience [1-4]. Systems for intelligent tutoring and other cutting-edge educational technology rely heavily on question generation [5-7]. This study presents a learning model for the second kind of question formats that leverage artificial intelligence using a machine learning model [6-9]. In the current research on ML approaches, ontologies are unable to provide grammatically acceptable questions [10]. Training computers to learn from text data without being explicitly programmed is the focus of machine learning, a subfield of computer science and artificial intelligence that is used in this research for automatic MCQ generation [11].

The algorithm is widely used in the universities for subject classification purposes. Classification and forecasting are common applications of the logit model, a form of statistical model. The dependent variable is restricted to values between 0 and 1 because the outcome is a probability. In logistic regression, the odds, defined as the likelihood of success divided by the probability of failure, are transformed using a logit function. This paper proposes a Semantic-based Enhanced Logistic Regression Model with Dual Mode Classification for accurate generation of MCQs that contains MCQs and odd one-model questions for the assessment of students. Teachers can save time by using automatic question generation to create test papers. M. Liu et al. [1] presented a mixed similarity approach that makes use of a statistical regression framework that incorporates orthographic, phonological, as well as semantic features. According to a recent evaluation of the educational QG market, there is a serious absence of easily accessible initial models for comparison as well as common baseline information that unifies various domains and question kinds. Hadifar et al. [2-6] aimed to bridge this gap with this paper.

2. Proposed Model

In this paper, an automated system is proposed for creating address labels that are quick, sleek, chaotic, and safe. This model is proposed to motivate efficient effort in producing programmed MCQ papers while requiring fewer people to do it because it needs expertise in both the

subject matter and the pedagogical processes involved, creating MCQs is a challenging assignment for humans. Since every sentence in the text can't be a candidate for MCQ, it's difficult for exam setters to build MCQs manually with an appropriate statement and relevant alternatives. Furthermore, it is a time-consuming effort to read the full topic and select significant lines or concepts for MCQ production. It is also difficult to be succinct without shifting the focus or eliminating relevant details while making decisions or stating a position. The stem, the key, and the distractor are the three components of MCQs. The question or statement being probed is the stem, the key, and the distractors, which may or may not include the right response, are the key and distractors, respectively. When the setter hasn't done their job well, it can take a long time for the student to read the stem and select the correct key from a pool of possible distractions. Therefore, it is necessary to show a system capable of creating the MCQs intelligently.

Distractors are created in this system by utilizing the lexical database WordNet, the internet resources, and the results of Google searches. A machine-readable dictionary is called a WordNet. It is an English language lexical database. Adjectives, adverbs, verbs, and nouns are all classified together into synonym sets in WordNet. Every word conveys a unique idea. These are interlinear synsets. Between the pieces in synsets, there are linguistic and semantic links. Similar to a thesaurus, WordNet offers the advantage of grouping words according to a specific sense. WordNet, for instance, created an analogous set of phrases associated with the concept of a key to lexicalize it. The basic MCQ generation module is shown in below figure 1.

The system operates in the following three key steps: extraction of informative sentences, identification of keys, and finally creation of distractors.

After preprocessing, the text file is delivered to feature processing so that it can create questions. Three steps go into creating the questions: selecting sentences, extracting key sentences, and selecting keywords. Sentence Selection is carried out by using words from the Topic_list, which are used to recognize the input text file's sentences. Topic_list is a manually compiled list by the subject specialist of pertinent and significant subjects for a domain used for MCQ generation. The professor who has completed the operating system course multiple times is the subject expert. The courses manually produced MCQs are selected, identified, and stored in the Topic_list by the subject matter expert. Sentence selection in the initial stage of MCQs is done using this list. If the topic word is absent from the

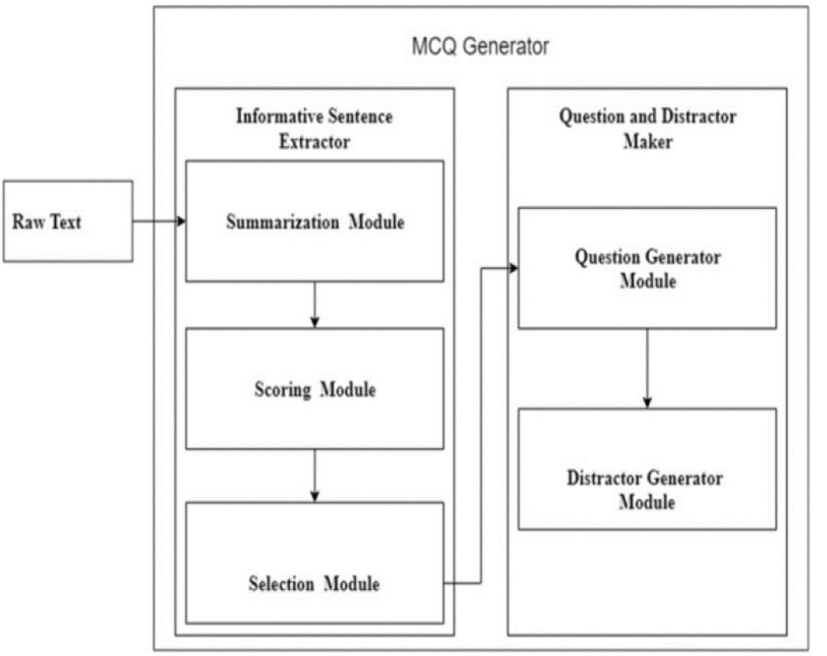

Figure 1

MCQ Generating Flow

input sentence, it is discarded. The first list of phrases sent to the subsequent phase is provided at this step.

Classifying and selecting the most important sentences from the phrases acquired in the first step constitutes the following stage of key phrase selection. In order to select the classifier that yields the greatest number of important sentences from the entire set of phrases, enhanced logistic regression is utilized in this process. The paper has adopted the highest number of crucial sentences as a criterion; similar to the education area, the goal is to provide the greatest number of pertinent questions. Features from the feature set are utilized to train the classifier. At this point, the sentences that meet at least one requirement are selected as key sentences.

Feature set consists of a number of characteristics like: Sentence Length (SL), sentences with a short length lack context, whereas sentences with an extended length would be too intricate to be selected for questions. The suggested study used 100–200 characters for the sequence length. More nouns in a phrase provide more contextual information. As a result,

the total amount of nouns in a phrase is determined using this attribute. A sentence that contains nouns is a suitable source for questions. Abbreviation is a boolean feature that determines if the text contains abbreviation. This characteristic determines if the sentence contains any abbreviation. As a result, this feature uses the existence or lack of abbreviations to judge the sentence's relevance. Key terms that are utilized as themes or subtopics to help find the appropriate sentences are included in the appendix at the conclusion of each book. To ascertain whether a sentence includes any Appendix Words (AW) that was taken from the appendix; this function's value is binary (T/F). These AW phrases are great starting points for creating questions. The proposed model framework is shown in below.

Initialize-Dataset→Pre-processing→Feature-Selection→Applying-Logistic-Regression-Analysis→Performance-Measure.

Pre-processing is performed as

$$\text{Valset}[M] = \sum_{r=1}^{M} \text{getattr}(r) + \max\text{range}(r) + \min\text{range}(r) + \lambda\left(\frac{\max(r,r+1)}{\min(r,r+1)}\right)$$

$$\text{V Norm}[M] = \sum_{r=1}^{M} \frac{\Sigma\max(\text{Valset}) - \min(\text{Valset})}{\text{len (Valset)}} + \text{getattr}(SS(r))$$

$$\text{VProcset}[M] = \sum_{r=1}^{M} \frac{\max(\text{VNorm}(r))}{10} + \frac{\text{VNorm}(r) - \lambda(\text{Valset}(r,r+1))}{\max(\text{VNorm})} - \min(\text{VNorm}(r,r+1))$$

Here λ is the model for mean calculation among the records. r is the current record and r+1 is the adjacent record. New features can be generated from existing subject features, or from a combination of existing features, for use in statistical analysis. The feature set generation is performed as

$$\text{Fset}[M] = \sum_{r=1}^{M} \frac{\max(\gamma(\text{VProcset}(r)))}{\text{VNorm}(r)} + \frac{\sum_{r=1}^{M}\text{simm}(\text{Feat}(r,r+1))}{M} + \text{corr}(\text{VProceset}(r,r+1)$$

Here γ is the model for considering the attributes that have maximum range of text data.

From the feature set generated from the subject material, keywords need to be identified that are used in MCQs generation.

$$\text{Keyword Set}[M] = \prod_{r=1}^{M} \max(\text{Fset}(r,r+1)) + \lim_{r\to M}\left(\max(\text{corr}(r,r+1))) + \frac{\text{maxrange}(Fset(i,i+1))}{M}\right)^2 - \min(\text{corr}(\text{Fset}(i,i+1))$$

Words and their interpretation adhere to a set of principles known as semantics. Semantics is the branch of linguistics concerned with meaning. The semantic rules processing is performed as

$$Sruleset[M] = \sum_{r=1}^{M} \max_{1 \le r \le M} \tau(\text{Keywords}) * \max \text{range}(Fset(r))$$

$$Sset[M] = \prod_{r=1}^{M} \frac{\sum \max(\tau(Fset(r))}{\min(corr(Fset))} + \mu(BoW(\text{KeywordsSet}(r, r+1)))$$

$$SRset[M] = \prod_{r=1}^{M} \frac{\sum \max(\text{Sset}(r, r+1))}{len(BoW(\text{Sruleset}(r))} + \max(\text{simm}(\text{Sset}(r, r+1))) - \min(\tau(\text{Sset}(r))$$

Here τ is the model for keyword frequency range. μ is the model for setting the rank for the keywords. In logistic regression, there is a single dependent variable that takes on the values 0 or 1, and the variables that provide explanation can be potentially discrete linear or continuous. The ELR model performs usage of ranking distinct features for better MCQs generation.

$$LR = \log\left\{\frac{\max(Fset(r, r+1))}{len(SRSet)}\right\} + max(corr(r, r+1)) + \max(\tau(SRset(r)))$$

$$LRset[M] = \sum_{r=1}^{M} LR(r) + (corr(r, r+1) * LR(r+1))$$

$$\log\left(\frac{Lrset(r)}{Lr(r)}\right) = \sum_{r=1}^{M} Srset(r) * \text{Keywordset}(r)$$

$$ELR[M] = \sqrt{\sum_{r=1}^{M} \max(LR(r, r+1)) + \sum_{r=1}^{M} \frac{\max(corr(r+1, \gamma(r) - \mu(r))}{mean(LRset(r))}} + \max(LRSet(r)) + \log\left(\frac{Lrset(r)}{Lr(r)}\right)$$

Keywords and words used in the classification could then be captured by a system. The dual mode classification is performed and the MCQs generation is performed as

$$CSset[M] = \sum_{r=1}^{M} \max(\tau(\text{KeywordSet}(r, r+1)) + \text{simm}(LRset(r, r+1)) + \frac{\max(corr(r+1, \gamma(r) - \mu(r))}{mean(LRset(r))} - \frac{\min(simm(r+1, \gamma(r) - \mu(r))}{\min(ELR(r, r+1))} \begin{cases} CSset \leftarrow SRset(r) & \text{if } simm(r, r+1) > Th \\ \text{continue} & \text{Otherwise} \end{cases}$$

$$Stem[M] = \frac{\sum_{r=1}^{=-34} \lambda(CSset(r))}{mean(ELR(r,r+1))} + \max(KeywordSet(r)) + BoW(KeywordsSet(r,r+1) + \frac{\max(\text{simm}(r+1,\gamma(r)-\mu(r))}{\min(diff(CSset(r))}$$

$$Key[M] = \prod_{r=1}^{M} \frac{\sum_{r=1} getattr(Stem(Keywordset(r,r+1)))}{max(corr(r,r+1))} + \max(corr(Stem(r), Keywordset(r)))$$

$$Distractor[M] = \prod_{r=1}^{M} \frac{\sum_{r=1} getattr(Stem(\max(\tau(SRset(r)))}{\max(corr(r+1,\gamma(r)-\mu(r)))} + \min(\text{diff}(corr(Stem(i), Keywordset(r)))) - \max(diff(corr(Stem(r), Keywordset(r))) + \frac{\sum \max(\tau(Fset(r))}{\min(corr(Fset))}$$

3. Results

Learning over the internet has grown popular and helps learners to learn anything, anytime, anywhere from the web resources. In any educational institution, assessment is crucial. Learners' ability to identify and fill up their learning gaps is greatly enhanced by an assessment system. It is a lot of work to generate questions manually. Therefore, automatic question production from learning resources is the core duty of an automated assessment system. Each university has its unique procedures for evaluating students subjectively. The proliferation of distance learning opportunities calls for the use of objective measures,

Table 1

MCQs Generation Accuracy Levels

Text Data Records Size	Models Considered		
	SbELR-DMC Model	EduQG Model	GMCQ-TCSLS
20000	97.4	91.6	90.7
40000	97.6	91.8	91.0
60000	97.8	92.0	91.2
80000	97.9	92.1	91.3
100000	98.1	92.3	91.5
120000	98.3	92.5	91.7

such as those provided by an automated evaluation system. Table 1 provide the MCQs generation accuracy levels details.

4. Conclusion

This article summarized and critically examined the methods of visual question production, as well as objective question generation with learner response evaluation. With the expansion of automation, we are passing over every duty to machines. The method generates MCQs based on the topic's imperative stems. A stem, key, and four distracters one correct answer, and three incorrect relevant answers make up each MCQ, six generally categorized dependent processes were identified: pre-processing, distractor production, question formation, sentence choice, key selection, and post-processing.

References

[1] M. Liu, V. Rus and L. Liu, "Automatic Chinese Multiple Choice Question Generation Using Mixed Similarity Strategy," *in IEEE Transactions on Learning Technologies*, vol. 11, no. 2, pp. 193-202, 1 April-June (2018), doi: 10.1109/TLT.2017.2679009.

[2] S. K. Hadifar, J. Bitew, C. Deleu, and T. Demeester, "EduQG: A multi-format multiple-choice dataset for the educational domain," *IEEE Access,* vol. 11, pp. 20885–20896 (2023), doi: 10.1109/ACCESS.2023.3248790.

[3] D. R. CH and S. K. Saha, "Generation of Multiple-Choice Questions from Textbook Contents of School-Level Subjects," *in IEEE Transactions on Learning Technologies*, vol. 16, no. 1, pp. 40-52, 1 Feb. (2023), doi: 10.1109/TLT.2022.3224232.

[4] J. Xie, N. Peng, Y. Cai, T. Wang and Q. Huang, "Diverse Distractor Generation for Constructing High-Quality Multiple-Choice Questions," *in IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 280-291 (2022), doi: 10.1109/TASLP.2021.3138706.

[5] M. K. Gupta, P. Dadheech, A. Kumar, S. R. Dogiwal, R. C. Poonia, L. Raja, and D. P. Bhatt, "Detection and localization for watermarking technique using LSB encryption for DICOM Image," *J. Discrete Math. Sci. Cryptogr.*, vol. 25, no. 1, pp. 193–204 (2022).

[6] M. Larrañaga, I. Aldabe, A. Arruarte, J. A. Elorriaga and M. Maritxalar, "A Qualitative Case Study on the Validation of Automatically

Generated Multiple-Choice Questions from Science Textbooks," *in IEEE Transactions on Learning Technologies*, vol. 15, no. 3, pp. 338-349, 1 June (2022), doi: 10.1109/TLT.2022.3171589.

[7] M. Aydoğan and A. Karci, "Improving the accuracy using pre-trained word embeddings on deep neural networks for turkish text classification", *Phys. A Stat. Mech. Appl.*, vol. 541, pp. 123288 (2020).

[8] G. Saini, V. Bhatnagar, S. Seema, L. Raja, S. Sharma, and R. C. Poonia, "Structural equation-based model to investigate the moderating effect of fear of COVID using partial least square method," *J. Interdiscip. Math.,* vol. 25, no. 3, pp. 703–720 (2022).

[9] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding", *Proc. Adv. Neural Inf. Process. Syst.*, pp. 5753-5763 (2019).

[10] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, and R. Kurzweil, "Multilingual universal sentence encoder for semantic retrieval," arXiv:1907.04307 (2019). [Online]. Available: http://arxiv.org/abs/1907.04307.

[11] T. Thongtan and T. Phienthrakul, "Sentiment classification using document embeddings trained with cosine similarity", *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics Student Res. Workshop*, pp. 407-414 (2019).

Received November, 2024