VideoDL: Video-Based Digital Learning Framework Using AI Question Generation and Answer Assessment

https://doi.org/10.3991/ijac.v16i1.35207

Abdur Rahim Mohammad Forkan[1(✉)], Yong-Bin Kang[1], Prem Prakash Jayaraman[1],
Hung Du[2], Steven Thomson[2], Elizabeth Kollias[2], Natalie Wieland[2]
[1]Swinburne University of Technology, Melbourne, Australia
[2]Vidversity, Melbourne, Australia
fforkan@swin.edu.au

Abstract—Assessing learners' understanding and competency in video-based digital learning is time-consuming and difficult for educators, as it requires the generation of accurate and valid questions from pre-recorded learning videos. This paper demonstrates VideoDL, a video-based learning framework powered by Artificial Intelligence (AI) that supports automatic question generation and answer assessment from videos. VideoDL comprises of various AI algorithms, and an interactive web-based user interface (UI) developed using the principles of human-centered design. Our empirical evaluation using real-world videos from multiple domains demonstrates the effectiveness of VideoDL.

Keywords—video-based learning, question generation, learning assessment, online learning

1 Introduction

The COVID-19 pandemic has forced education sectors to adopt digital learning in particular the asynchronous mode of teaching using education videos (i.e., learners watch pre-recorded lecture videos at their own pace). This demands a reliable framework to quickly assess learners' understanding and competency based on provided digital contents of videos. To assess learners' competency for a given learning video, teachers need to go through the whole contents and manually form a set of relevant questions along with answers. Moreover, teachers have to manually assess learner-provided answers to those questions to complete assessments. Assessments can vary in the context of the student cohort and relevant band or level of learning, thus such efforts are time-consuming, inefficient and arduous.

Artificial intelligence (AI) can promote interactive communications between teachers and learners in a digital learning environment [6]. AI techniques are now extensively used to reduce manual effort by teachers. Research progress has been made in Automatic Question Generation (AQG) from textual data based on syntax and semantics [5]. For instance, researchers used deep learning techniques for AQG such as BERT [2], T5 transformer language model [8], GPT-2, and GPT-3 language model [11]. Moreover, a text-based similarity measure such as sentence-BERT (SBERT) [12] was used for automatic answer assessment (AAA) or grading [10]. The paradigm of AQG and AAA can make teachers more efficient. Therefore, AQG can have capability to generate various types of quality questions that teachers desire to utilize for assessment, and AAA can be used as an essential supporting tool for grading.

This paper demonstrates VideoDL, a video-based learning framework, developed in collaboration with VidVersity [14], an Australian company promoting video-based education. VideoDL incorporates recent AI algorithms for generating four different types of questions from pre-recorded video lectures and perform automatic assessment of answers. VideoDL takes a human-centered approach wherein teachers can optionally modify/edit AI-generated questions and AI-recommended answer assessments using an interactive UI.

2 VideoDL design

The key features of VideoDL comprising of Question Generation Platform (QG-P) and Learning Assessment Platform (LA-P) are as follows (see Figure 1). First, VideoDL takes a user-centered design approach, optionally enabling teachers to interact with the AI modules to refine the AI-generated questions (see the components of 'Teacher Involvement' in Figure 1). VideoDL has been co-designed with educators who bring significant experience in delivering digital learning and teaching outcomes. Second, QG-P performs a 5-phase pipeline to generate the four types of questions (i.e., short-answer, Boolean, gap fill, and multiple-choice question types – abbreviated by SAQ, BLQ, GFO and MCQ, respectively) from a given video. As discussed in ref. [5], most existing works paid attention to generate objective type questions (e.g., MCQs or BLQs), and recently more research works are interested in generating subjective type questions (e.g., SAQs and GFQs). VideoDL incorporates various AI techniques into QG-P to generate both objective and subjective types of questions. Third, LA-P has been designed to assess learner understanding and proficiency on learning materials on the video. Depending on the question type, LA-P uses a different assessment metric.

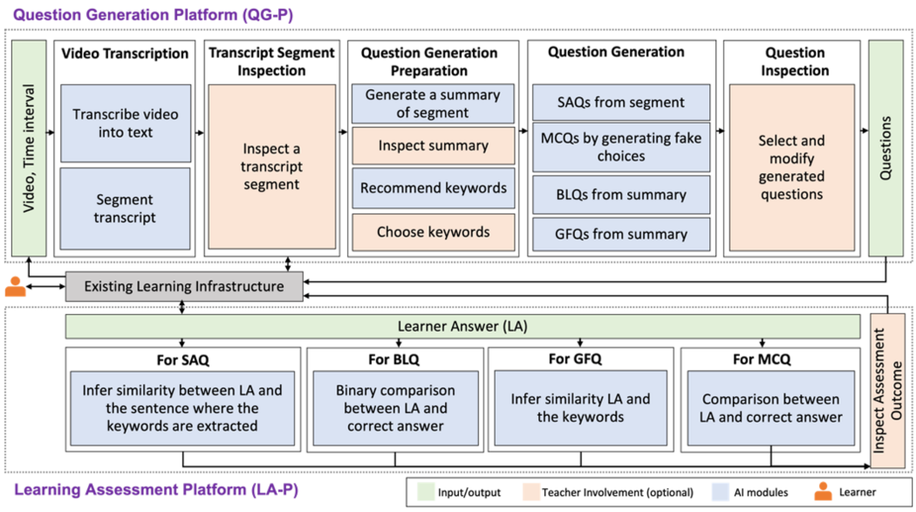

Fig. 1. VideoDL comprising 'Question Generation Platform (QG-P)' and 'Learning Assessment Platform (LA-P)'

2.1 Question Generation Platform: QG-P

In 'Video Transcription,' QG-P transcribes a given video v to text (or transcript) t_v using a speech recognition technique. Then, t_v is divided into n segments $S^{tv} = \{s_1, \ldots, s_n\}$. A segment-based education has already been shown effective in education settings [1]. Thus, QG-P generates questions from a segment $s \in S^{tv}$. The segmentation can be done by a time duration (e.g., minutes) given by the educator. In 'Transcript Segment Inspection,' given s, QG-P optionally allows the educator to inspect s to see if there are some errors or noise text to be fixed. If necessary, s is fixed, and the updated segment is stored in the existing learning infrastructure.

In 'Question Generation Preparation,' QG-P generates three kinds of data essentially used for question generation in the next phase. Given s (or updated s in the prior phase), (a) it generates s's *abstractive* summary (aiming to automatically generate a smaller and concise piece of s) as_s to be used as the input to generate BLQs and GFQs. The summary generation is performed using a pretrained language model, Google's T5 (Text-to-Text Transfer Transformer) [13]. Once as_s is generated, the educator can also manually inspect it (optional), and if necessary, he/she can update as_s to improve language fluency. (b) SAQs and MCQs are created based on target key concepts (which we term as *keywords*) that appear in s. A keyword is a span of text from s around which a question is generated. QG-P automatically recommends top-N (N is a parameter) candidates of keywords (both noun phrases and named entities) from s. Optionally, the educator can select keywords from the candidates or additionally choose some words manually as keywords. Furthermore, (c) QG-P also generates a set of keywords from as_s for generating GFQs, optionally interacting with the educator as (b), that is, the educator can choose keywords manually or from the recommendation of QG-P. In a GFQ, one or more words are removed from as_s, and this incomplete text is given to a leaner as a question.

In 'Question Generation,' QG-P generates SAQs from s and the chosen keywords K = $\{k_1, \ldots, k_n\}$. For this, QG-P uses ParaQG [9] that can generate fluent, relevant questions from s and each k_i (seen as the correct answer). Also, QG-P generates BLQs that can go beyond what is immediately stated in s. BLQs can thus be used to assess an overall comprehension of the leaner about key information delivered from s. QG-P generates BLQs from the summary ass using the T5-base model [13] trained on Bool [4]. Also, this model can generate the correct answers for the generated questions. Moreover, QG-P generates GFQs from the summary as_s, where a GFQ is a question that the learner is asked to fill one or more omitted words (i.e., the correct answers) given a text. Finally, QG-P generates MCQs from s to assess specific knowledge embedded in s. For this, QG-P uses SAQs along with a distractor generation model [3] that can generate multiple context-related incorrect choices.

Finally, in 'Question Inspection,' the educator can optionally inspect AI-generated questions. If necessary, they can manually modify. The updated questions are finally stored in the existing learning infrastructure.

2.2 Learning Assessing Platform: LA-P

LA-P has been designed to help the educator to assess a learner-provided answer x given a question q using the correct answer z. Given a BLQ, assessment is straightforward by comparing the x with z, where both answers are given as yes/no or true/false. Also, MCQs can be simply assessed by comparing the correct answer choice that is the keyword and x. Given a SAQ q, LA-P uses the SBERT model [12] to infer a contextual similarity between x and z.

LA-P incorporates two similarity functions, and our evaluation shows which one performs better. The first measures similarity sim between two text snippets x and z using SBERT denoted as $\mathbf{SIM}_{base}(x, z)$. The other function measures a semantic similarity between x and z, considering the context where z was extracted, denoted as: $\mathbf{SIM}_{sent}(x, \text{context}[z])$, where context($z$) is the context of z. As such a context, we use the 'sentence' that encompasses z. Sentence is generally seen as a linguistic unit consisting of words that are meaningfully linked together [7] By exploiting the context, we aim to enhance $\mathbf{SIM}_{base}(x, z)$. Assessing x to a GFQ is relatively simple by examining whether x is close to z. We use $\mathbf{SIM}_{base}(x, z)$ to measure their similarity.

3 Demonstration and evaluation

A snapshot of the VideoDL UI is presented in Figure 2 that implements the VideoDL's 5-phases for QG-P, and easy-to-use steps for LA-P. Figure 2a shows a ranked list of recommended keywords (purple), and manually chosen keywords by the teacher (green). As discussed, SAQs and MCQs will be generated based on each of the chosen keywords at this phase. Figure 2b shows an example of a SAQ generated based on the given keyword, 'report ill health' (in this example, we generated top-3 SAQs). Figure 2c shows examples of a GFQ and how a learner's answer is assessed by our similarity measure, $\mathbf{SIM}_{base}$. The GFQ (the left image) was derived from the summary of the original transcript in Figure 2b, which was generated using Google's T5 model as described in Section 2 (see the right image). The middle image shows the similarity scores between the learner's answers and the correct answers. Based on the similarity scores, the teach can make final assessment. A demonstration video of VideoDL is also available at https://youtu.be/c8IiYtu9Gjs.

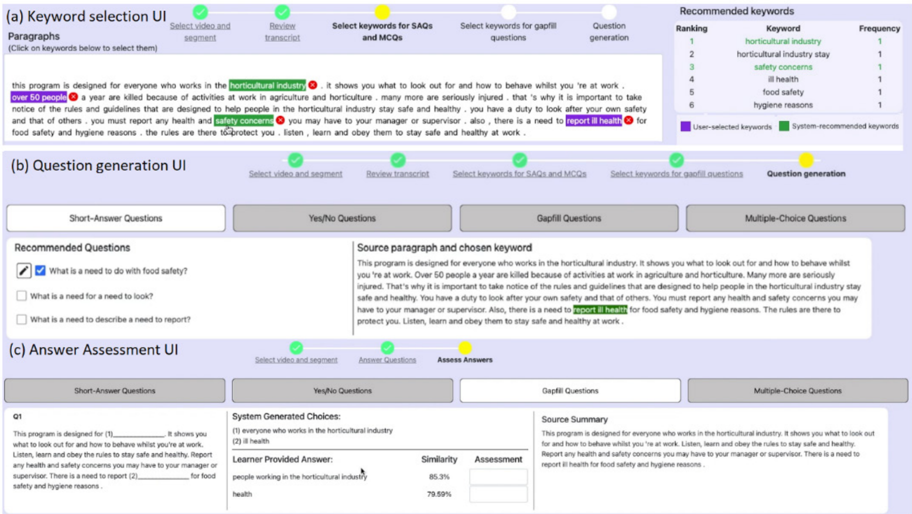

Fig. 2. A snapshot of the interactive VideoDL UI. (a) Keyword selection (part of Question Generation Preparation in QG-P). (b) Question Inspection in QG-P. (c) Generated similarity (or assessment) scores by LA-P between learner-provided answers and the correct answer

To evaluate the quality of four types of generated questions (SAQ, BLQ, MCQ and GFQ) by QG-P, we used 117 educational videos from 12 different education domains (e.g., law, banking, finance, leadership). The lengths of videos varied from 30 seconds to 1 hour. The average length of generated video segments was 4 minutes 35 seconds. Seven experienced teaching professionals assessed the quality of generated questions by rating them into 3 categories: "Good," "Average," and "Bad." In the question-generation process, 23% recommended keywords were used, and the remaining 77% were manually chosen by the professionals. Table 1 shows the evaluation outcome. As observed, acceptable questions ("Good" and "Average" rated) were dominant in all the 4 types of questions.

Table 1. Evaluation of generated questions by VideoDL

Type	Question No.	Good	Average	Bad
SAQ	335	39%	33%	28%
BLQ	164	40%	26%	34%
MCQ	346	51%	27%	22%
GFQ	116	85%	12%	3%

To examine possible reasons for the "Bad" rated questions in SAQs and BLQs, we conducted qualitative analysis on the segments and keywords used to generate those questions (analysis on MCQs is our future work, and GFQs are excluded as their "Bad" ratings reach only 3%). The analysis results are presented in Table 2 that guides us how we can further enhance VideoDL's capability. The dominant reason was incorrect choices of keywords by the professionals. To address the issue, we may further need to identify what are the evaluators' rationale for choosing such incorrect keywords and incorporate their approach for choosing good keywords for question generation into. The other reasons, except for "unknown," were identified as incompleteness of the AI modules. This indicates AI for question generation still has a room for further improvement, despite VideoDL has been equipped with state-of-the-art AI techniques. We believe that this fact drives which areas we need to work more on in the future research.

Table 2. Analysis on bad rated questions on SAQs and BLQs

Reason	SAQs	BLQs
Bad Keyword selection by the professionals	52%	–
Weakly associated questions generated in QG-P	13%	53%
Noise in text when transcribing video in QG-P	29%	25%
Wrong recommendation of answer (Yes/No) in QG-P	–	18%
Reasons unknown	6%	4%

To evaluate LA-P, we used SAQs as this type of questions is viewed as the most difficult subjective questions that require longer time to assess by teachers. A total of 129 SAQs generated by QG-P was randomly selected. The same 7 teaching professionals were asked to (a) provide answers to those SAQs from learners' perspectives and (b) grade the answers using an assessment score (from 0 to 100), t_s. Then, we measured $\mathbf{SIM}_{base}$ and $\mathbf{SIM}_{sent}$ between the correct answers (the keywords chosen when generating questions) and the provided answers. Finally, we measured Pearson correlation coefficient ρ between the similarity scores for each similarity method and t_s for the 129 SAQs. The $\rho(\mathbf{SIM}_{base}, t_s)$ was 0.423 and $\rho(\mathbf{SIM}_{sent}, t_s)$ was 0.497. It indicates that $\mathbf{SIM}_{sent}$ is closer than $\mathbf{SIM}_{base}$ to the human judgement of the evaluators, and the positive ρ values justify the validity of our similarity measures. Here, a higher ρ means our similarity score is closer to t_s. With this sample set of SAQs, we observed that it is still challenging to achieve stronger human-level assessment (stronger positive correlation) and this requires further research.

4 Conclusion

This paper presents a demonstration of VideoDL that has been developed for human-centered question generation and answer assessment from educational videos. VideoDL incorporates recent AI techniques with teachers' knowledge to generate reliable, practical questions. Moreover, VideoDL is also designed to help teachers to facilitate answer assessment. We demonstrate VideoDL's functionalities through a web-based UI. Furthermore, our evaluation shows the practicability and effectiveness of VideoDL using more than 100 videos from 12 educational domains.

5 References

[1] Cynthia J Brame. 2016. Effective educational videos: Principles and guidelines for maximizing student learning from video content. CBE Life Sciences Education, 15(4):es6. https://doi.org/10.1187/cbe.16-03-0125

[2] Dhawaleswar Rao Ch and Sujan Kumar Saha. 2018. Automatic multiple choice question generation from text: A survey. IEEE Transactions on Learning Technologies, 13(1):14–25. https://doi.org/10.1109/TLT.2018.2889100

[3] Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2020. A bert-based distractor generation scheme with multi-tasking and negative answer training strategies. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pages 4390–4400. https://doi.org/10.18653/v1/2020.findings-emnlp.393

[4] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. arXiv preprint arXiv:1905.10044.

[5] Bidyut Das, Mukta Majumder, Santanu Phadikar, and Arif Ahmed Sekh. 2021. Automatic question generation and answer assessment: A survey. Research and Practice in Technology Enhanced Learning, 16(1):1–15. https://doi.org/10.1186/s41039-021-00151-1

[6] Brett D. Jones. 2020. Motivating and engaging students using educational technologies. In Handbook of Research in Educational Communications and Technology: Learning Design, pages 9–35. Springer International Publishing. https://doi.org/10.1007/978-3-030-36119-8_2

[7] Yong-Bin Kang, Pari Delir Haghigh, and Frada Burstein. 2016. Taxonomy. IEEE Transactions on Knowledge and Data Engineering, 28(2):524–536. https://doi.org/10.1109/TKDE.2015.2475759

[8] Parteek Kumar. 2021. Deep learning based question generation using t5 transformer. In Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I, volume 1367, page 243. Springer Nature. https://doi.org/10.1007/978-981-16-0401-0_18

[9] Vishwajeet Kumar, Sivaanandh Muneeswaran, Ganesh Ramakrishnan, and Yuan-Fang Li. 2019. Paraqg: A system for generating questions and answers from paragraphs. EMNLP-IJCNLP 2019, page 175. https://doi.org/10.18653/v1/D19-3030

[10] Ifeanyi G Ndukwe, Chukwudi E Amadi, Larian M Nkomo, and Ben K Daniel. 2020. Automatic grading system using sentence-bert network. In International Conference on Artificial Intelligence in Education, pages 224–227. Springer. https://doi.org/10.1007/978-3-030-52240-7_41

[11] David Oniani and Yanshan Wang. 2020. A qualitative evaluation of language models on automatic question-answering for covid-19. In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pages 1–9. https://doi.org/10.1145/3388440.3412413

[12] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

[13] A Roberts and C Raffel. 2020. Exploring transfer learning with t5: The text-to-text transfer transformer. Google AI Blog.

[14] Vidversity. 2022. https://vidversity.com/. Last accessed 14 February 2022.

6 Authors

Abdur Rahim Mohammad Forkan, Senior Research Fellow, AI and Machine Learning in the Swinburne Digital Innovation Lab. He has expert knowledge on machine learning with significant experience developing AI-based solutions in digital health, FinTech, Agriculture, Education and Industry 4.0 application areas. He obtained his PhD in Computer Science from RMIT University, Australia. His research interests include data science, health informatics, pervasive computing, and applied AI (email: fforkan@swin.edu.au).

Yong-Bin Kang was awarded a PhD in Faculty of IT from Monash University and is now a senior data science research fellow for the ARC Centre of Excellence for Automated Decision Making and Society (ADM+S) at Swinburne University of Technology. His research expertise is primarily focused on knowledge discovery, predictive modelling and decision-making optimization through advanced natural language processing, machine learning, and AI techniques. Recently, his research has focused more on developing, managing, and delivering algorithms and solutions that improve open, ethical, and trustworthy AI analytics practices for the social good (email: ykang@swin.edu.au).

Prem Prakash Jayaraman is an Associate Professor and Director of the Factory of Future and Digital Innovation Lab at Swinburne University of Technology. He leads industry-funded ground-breaking research in Internet of Things (IoT), Mobile and Cloud computing to cocreate novel solutions underpinned by digital technologies solving industry problems and aiding their digital transformation journey. He has authored or co-authored 140+ journal articles, conference writings, and book chapters in highly ranked venues the above research areas. (email: pjayaraman@swin.edu.au).

Hung Du is a research engineer at Applied Artificial Intelligence Institute (A2I2), Australia. He has experience in translating research into real-world applications in a wide range of industrial domains such as finance, education, compliance, automation, and software engineering. He received the ACM SIGSOFT Distinguished Paper Awards at IEEE/ACM 19th International Conference on Mining Software Repositories. His research interested include Natural Language Processing, Reinforcement Learning, Machine Learning Operations and Applied Artificial Intelligence. (email: peter@vidversity.com).

Steven Thomson has been working in the IT industry for over ten years as an IT administrator for businesses and consumers. More recently, he has been studying at Swinburne university under a Bachelor of Computer Science and work in the software development field. (email: steven@vidversity.com).

Elizabeth Kollias started her professional career as a lawyer, and quickly took a keen interest in the management of the firm. This combined with her passion for writing, led to a role at CPD Interactive in customer development and communications. She moved into the digital learning creation space, with a focus on learning design, communication and strategy generally. She originally came on board as a the COO but with her love of detail and vision she moved into the CEO role. (email: liz@vidversity.com).

Natalie Wieland began her working life as a lawyer in the mid 1990s, but soon went to study Management Information Systems. In addition to a sessional teaching role at the University of Melbourne and her ongoing consulting work, she is co-founder of VidVersity which is an all in one Australian solution for the creation and delivery of online training. She brings her real-world experience of teaching and training to the creation of online learning and training which she believes should be accessible to everyone (email: natalie@vidversity.com).

Article submitted 2022-09-07. Resubmitted 2022-11-01. Final acceptance 2022-11-10. Final version published as submitted by the authors.