Ontologies for Automatic Question Generation

by

Noor Hasimah binti Ibrahim Teo

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Computer Science

University of Warwick, Department of Computer Science

October 2019

Table of Contents

List of Figures

List of Tables

Acknowledgement

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Dr Mike Joy for the continuous support of my PhD study and related research, for his patience, motivation, and immense knowledge. His guidance helped me during the whole period of the research and writing of this thesis.

Besides my supervisor, I would like to thank my advisors: Dr Jane Sinclair, Prof. Dr Alexandra I. Cristea, and Dr Steve Matthews, for their insightful comments and encouragement, but also for the hard questions which incentivises me to widen my research from various perspectives.

I also thank my fellow labmates for the stimulating discussion, for the motivational advice, and for the days we were working together before deadlines, and for all the fun we have had in the last four years.

I am grateful to my beloved dad Ibrahim Teo and my dear mom Hasiah, who supported me and helped me throughout my life and during this study. Dad, thank you for all the sacrifices that you've made on my behalf. You are the reason for me to pursue this study, and I try my very best to make your dreams come true. Papa, this is for you.

I owe thanks to a very special person, my husband, Zuleaizal for his continued and unfailing love, unconditional support and understanding during this study that made the completion of this thesis possible. I greatly value his contribution and deeply appreciate his belief in me.

Last but not least, to my wonderful children, Mikail, Sofea and Zayn, thank you for bearing with me and my mood swings and being my greatest supporters.

Declaration

This thesis is presented by following the regulations for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work in this thesis has been undertaken by myself under the supervision of Dr. Mike Joy. Some parts of this thesis are written based on previously published papers (as the first author). Detail of all publications is described below.

- The validation of Course ontology element analysis discussed in Chapter 3 is partially based on the result obtained from experiments results and literature review published in:
    - **Ibrahim Teo**, **N.H.** and Joy, M.(2016). "Validation of Course Ontology Elements for Automatic Question Generation." 3rd EAI International Conference on e-Learning e-Education and Online Training (eLEOT).

- The design of question generation strategies and error analysis of the inappropriate representation of ontology concepts for assessment question keywords discussed in Chapter 4 is a further revision of work published in:
    - **Ibrahim Teo**, **N.H.** and Joy, M.(2016). "Evaluation of Automatic Question Generation Approach using Ontology." 15th European Conference on eLearning (ECeL).
      Award: *Certificate of Merit for PhD Paper and presentation*

- The design of categorised question templates discussed in Chapter 4 is based on the paper published in:
    - **Ibrahim Teo**, **N.H.** and Joy, M.(2018). "Categorized Question Template Generation for Ontology-based Assessment

Questions." The 7th International Conference on Knowledge and Education Technology (ICKET).

- In addition, the conference paper presented at eLEOT 2016, was selected to be extended and included in a journal of EAI Endorsed Transaction on e-Learning.
  - **Ibrahim Teo**, **N.H.** and Joy, M (2017). Validation of Course Ontology Elements for Automatic Question Generation. EAI Endorsed Transaction on e-Learning. 17(15).

- Besides, there are two poster presentation attended.
  - **Ibrahim Teo**, **N.H.** and Joy, M (2016). Automatic Question Generation from Ontology. Poster presented at Research Poster Competition, University of Warwick.

  - **Ibrahim Teo**, **N.H.** and Joy, M (2016). Evaluation of Automatic Question Generation Approach using Ontology. Poster Presented at 15th European Conference on eLearning (ECeL).

Abstract

Assessment is an important tool for formal learning, especially in higher education. At present, many universities use online assessment systems where questions are entered manually into a question bank system. This kind of system requires the instructor's time and effort to construct questions manually. The main aim of this thesis is, therefore, to contribute to the investigation of new question generation strategies for short/long answer questions in order to allow for the development of automatic factual question generation from an ontology for educational assessment purposes. This research is guided by four research questions: (1) How well can an ontology be used for generating factual assessment questions? (2) How can questions be generated from course ontology? (3) Are the ontological question generation strategies able to generate acceptable assessment questions? and (4) Do the topic-based indexing able to improve the feasibility of AQGen.

We firstly conduct ontology validation to evaluate the appropriateness of concept representation using a competency question approach. We used revision questions from the textbook to obtain keyword (in revision questions) and a concept (in the ontology) matching. The results show that only half of the ontology concepts matched the keywords. We took further investigation on the unmatched concepts and found some incorrect concept naming and later suggest a guideline for an appropriate concept naming. At the same time, we introduce validation of ontology using revision questions as competency questions to check for ontology completeness. Furthermore, we also proposed 17 short/long answer question templates for 3 question categories, namely definition, concept completion and comparison.

In the subsequent part of the thesis, we develop the AQGen tool and evaluate the generated questions. Two Computer Science subjects, namely OS and CNS, are chosen to evaluate AQGen generated questions. We conduct a questionnaire survey from 17 domain experts to identify experts' agreement on the acceptability measure of AQGen generated questions. The experts' agreements for acceptability measure are favourable, and it is reported that three of the four QG strategies proposed can generate acceptable questions. It has generated thousands of questions from the 3 question categories. AQGen is updated with question selection to generate a feasible question set from a tremendous amount of generated questions before. We have suggested topic-based indexing with the purpose to assert knowledge about topic chapters into ontology representation for question selection. The topic indexing shows a feasible result for filtering question by topics.

Finally, our results contribute to an understanding of ontology element representation for question generations and how to automatically generate questions from ontology for education assessment.

Abbreviations

AC	Agreement Coefficient
AI	Artificial Intelligence
AQGen	Automatic Question Generator
CNS	Computer Network & Security
DSC	Dice’s Similarity Coefficient
FIB	Fill-In-the-Blank
MCQ	Multiple Choice Question
MOODLE	Modular Object-Oriented Dynamic Learning Environment
NLP	Natural Language Processing
ontoCN	Computer Network & Security Ontology
ontoOS	Operating System Ontology
OS	Operating System
QD	Question Deficiencies
QG	Question Generation
QGS	Question Generation Strategy
S/LAQ	Short/long answer question
T / F	True / False

Chapter 1

Introduction

1.1 Background

An assessment has a major impact on learning as it provides data for further evaluation to allow benchmarks to be established for education standards. The overall process of assessment involves gathering data, interpreting, and making appropriate reporting. Assessment activities such as setting, responding to and marking the assessment are expensive and consume significant amounts of time and effort for both students and instructors. Assessment is important for many reasons, such as to provide feedback on learning and teaching to both the student and the teacher, allow students to be graded or ranked and to encourage learning (Broadfoot & Black, 2004). Apart from demonstrating a student's level of knowledge, assessments determine how students engage with the learnt material (Biggs, 2002).

Questions form an important mechanism in assessment for assessing student's knowledge. They are commonly used as a metric to determine the level of knowledge of students. However, preparation of assessment questions is cumbersome and challenging, especially for lecturers teaching new subjects. The problem becomes significant among novice lecturers. Usually, lecturers prepare complete sets of question papers manually and store them in a question bank. Therefore, we work toward automating the process of creating assessment questions.

Many research projects have looked at automating the process of creating assessment questions. Research in question generation began in the early 70s using natural language processing techniques as well as statistical methods. These techniques demand many complex rules and language processing. Recent research has led to the emergence of ontology-based question generation. The reasons for question generation include formative and summative assessment (Papasalouros et al., 2008; Mitkov et al., 2006), exercise questions (Lin & Chou, 2011), problem solving questions (Kumar, 2005), domain-specific general questions (Soldatova & Mizoguchi, 2009), or general questions (Hovy et al., 2001).

Research on ontology-based question generation has explored different question types which include multiple-choice questions (MCQ), verification, short answers and fill-in-the-blank. Different ontological question generation strategies have been explored and may be categorised into three main strategies which are class-based, terminology-based strategies and property-based strategies (Papasalouros et al. 2008; Cubric & Tosic, 2011; Al-Yahya 2014). They implement strategies to create what, when, why, where, and how questions for MCQ types of question., Although various automated question generator approaches have been studied, to the best of our knowledge, generating short/long answer questions of a definition, concept completion and comparison types have not yet been explored.

The overall motivation of this research is to be able to construct assessment questions by using a course ontology and evaluate the acceptability measure for generated questions against five question deficiencies. The focus of the work has led to the design of question generation strategies and development of question generation tools to evaluate the acceptability measure for the generated questions.

The remaining sections of this chapter present an illustrative example of questions generated using an ontology as a source, followed by the research methodology and details of the thesis structure.

1.2 Illustrative Example of Question Generation from an Ontology

This section provides an illustrative example of question generation using an ontology. We begin with a simple example to show how this idea will work.

Figure 1.1 is part of the concept representation of the Operating System subject in ontology. A simple question can be generated using any concepts in this ontology, such as:

i. Define virtual memory?

ii. Differentiate between swapping and segmentation.

iii. Explain the term demand paging in memory management.

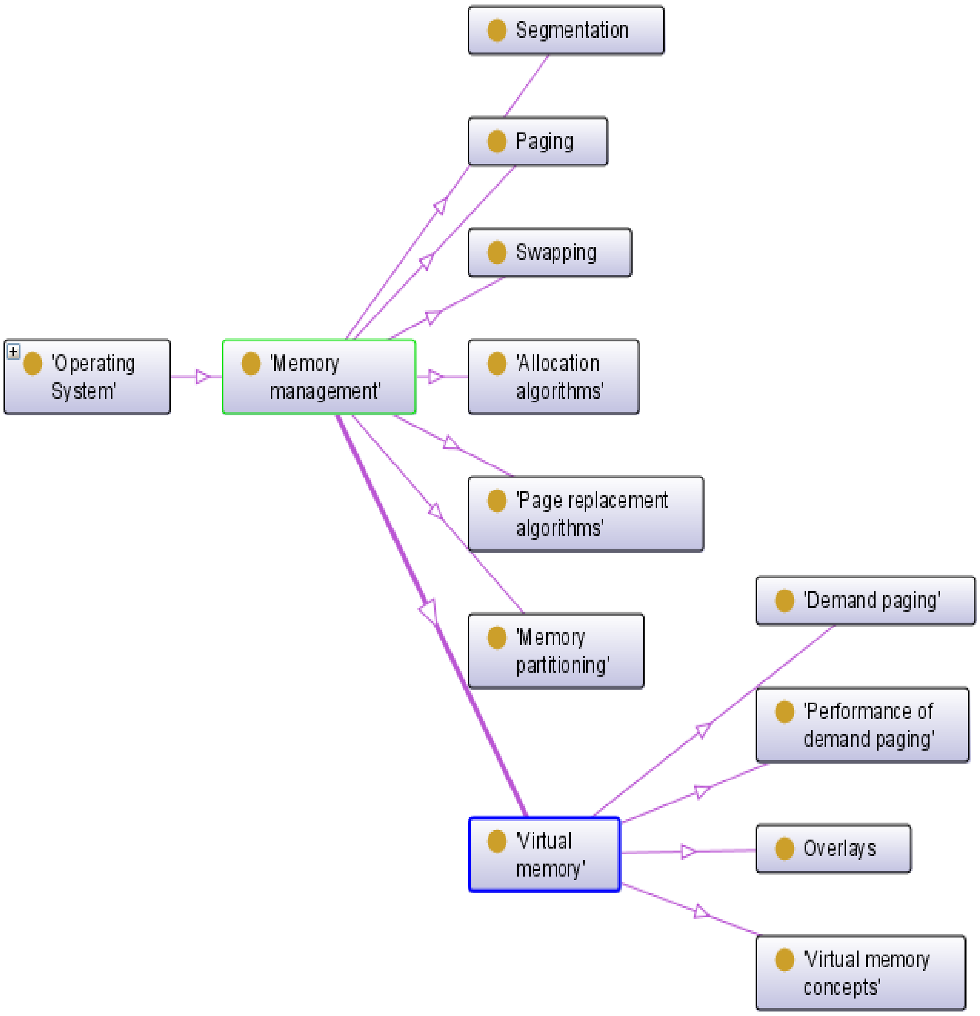


Figure 1.1: Part of Operating System Ontology representation.

Although keywords can be extracted from the ontology concepts, not all question templates are suitable for each of the concepts, especially for a question that contains two or more keywords. For example, the question

"Explain overlay in swapping" might not make sense since overlay only happens in virtual memory. Another example is the comparison type of question, "Differentiate between swapping and interrupt handling". Although this question is grammatically and syntactically correct, the comparison type of question usually differentiates between two keywords that fall into the same topic. Different question structures need different strategies to map question templates with the concepts in the ontology.

Therefore, this work will focus on strategies to generate acceptable questions that use the concepts that exist in the ontology as keywords for the questions.

1.3 Research Objective and Research Questions (RQ)

The main objective of this research is to contribute to the investigation of new question generation strategies in order to allow for the development of automatic factual and descriptive question generation from an ontology for educational assessment purposes. The assessment in this work refers to questions that are used to assess student knowledge, which include tests, quizzes and revision questions.

The main research questions which have guided this research work are as follows.

RQ1: How well can an ontology be used for generating factual assessment questions?

> RQ1.1: Do the course ontologies evaluated in this research contain concepts that are appropriate for the subject being evaluated?
>
> RQ1.2: What are the categories of question that can be generated using the ontology?

RQ2: How can questions be generated from course ontology?

> RQ2.1: How should categorized question templates be constructed?

RQ2.2: What ontological question generation strategies can be applied to generate questions?

RQ3: Are the ontological question generation strategies able to generate acceptable assessment questions?

RQ3.1: What is the agreement between experts for each QG strategy?

RQ3.2: To what extent do subject matter expert agree with the acceptability measure of generated questions?

RQ3.3: Are the generated questions grammatically correct?

RQ3.4: Are the generated question understandable?

RQ3.5: Are the generated questions relevant to the subject being evaluated?

RQ3.6: Can the generated questions be used as assessment questions?

RQ3.7: Do the generated questions use the correct question word?

RQ3.8: Is there a difference in QDs rating between the question generation strategies?

RQ3.9: Is there a difference in QD rating between questions generated using ontoOS and ontoCN?

RQ4: Do the topic-based indexing able to improve the feasibility of AQGen?

1.4 Research Methodology

This section discusses the research methodology adopted in this research as well as the research methods undertaken to address the research questions.

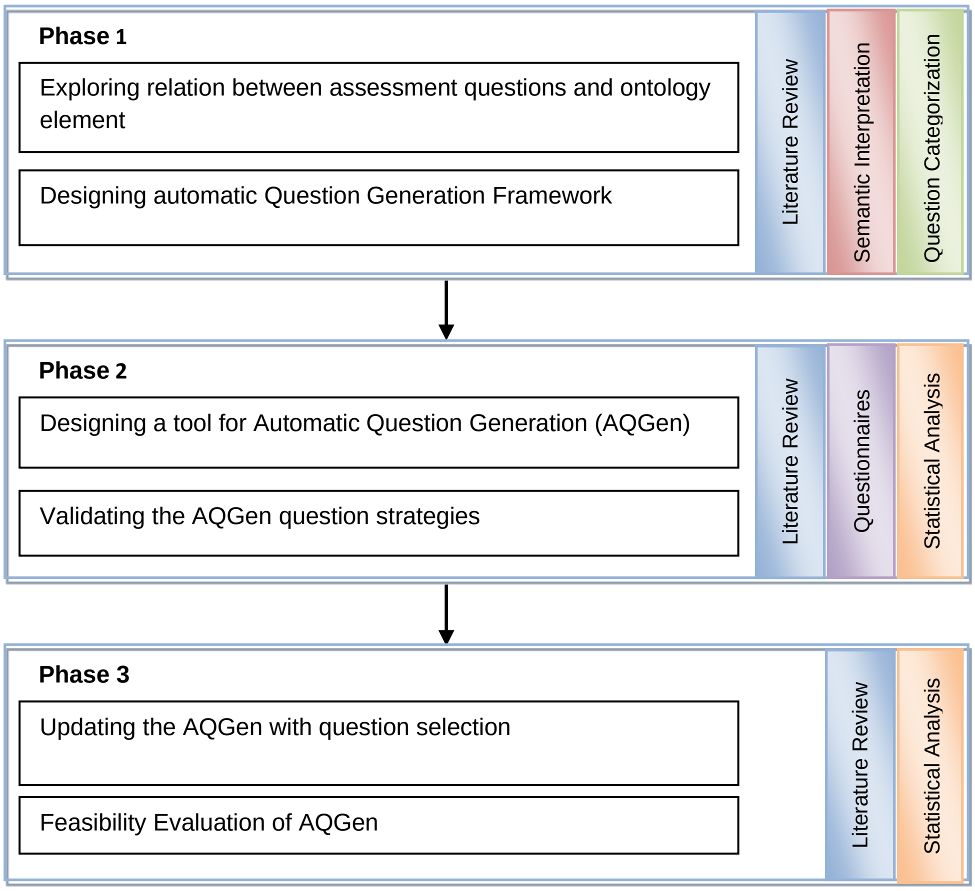


Figure 1.2: Three main phases of the research methodology. Components in the first column represent activities conducted, while components in the second column represent methods used in the research. Solid arrows indicate flows in the experimental process.

An inductive research approach was adopted in this work to propose new question generation strategies for factual assessment questions. The proposed Question Generation (QG) strategies are based on the underlying theory of cognitive computational model of question asking (Graesser & Olde, 2003). The inductive approach was applied in this work to evaluate existing ontological approaches for question generation using different question categories and ontologies. In order to initiate this work, four main research questions were outlined in the previous section, forming guidance throughout this study. The research questions are answered in three different phases of the research, as shown in Figure 1.2

1.4.1 Phase 1

There are two main research questions answered in this phase in order to explore the relationship between assessment questions and ontology elements. The work begins by answering RQ1, to identify how well course ontology from the publicly available ontology library on the web can be used as a source for a question generator. For the purpose of this evaluation, a course ontology for Operating Systems from the computer science domain has been used, due to the better understanding of the researcher in this field and the availability of the ontology online. In order to answer the main research question RQ1, a pre-analysis was undertaken to answer sub-research questions RQ1.1 and RQ1.2. These were formulated to identify the appropriateness of the ontology as well as suitable question types that can be generated using the ontology.

RQ1.1 identifies whether the concepts in the ontology represent the actual keywords assessed in the Operating System subject. An experiment was conducted to measure the string similarity between keywords existing in the assessment questions with the concepts inside the ontology, using the data mining technique consisting of n-gram models and dice coefficient similarity. The results indicate that the ontology was appropriate for the question generation purposes with more than half of the concepts in the ontology similar to keywords used in assessment questions. On the other hand, RQ1.2 sought to identify the question type that could be generated using the ontology. Existing revision questions from an Operating System textbook and past years examination questions were analysed, and the selection of questions was based on the predefined question structure. The analysis revealed different question categories that could be the basis for formulating question templates later.

The next step analysed how to generate questions from the ontology in RQ2. The results obtained from RQ1 were used as a sample to design question templates, which is the aim of RQ2.1. A series of steps to create questions templates are explained in Chapter 4 of this thesis. Moreover, RQ2.2

investigates question generation strategies using ontological approaches. Four question generation strategies were identified based on the question categories identified in RQ1.2.

Finally, based on the literature review and the results obtained, a conceptual framework for automatic question generation strategies was developed. The framework explains the process and mechanism used in automatic question generation.

1.4.2 Phase 2

RQ3 evaluates the ontological question generation strategies. A question generator tool was developed using the Java platform to generate assessment questions based on question templates produced in RQ2.1 and a set of question generation strategies proposed in RQ2.2. A list of generated questions was randomly sampled, and questionnaires were used to collect feedback from raters. Raters in this work refer to an expert, who is the instructor or postgraduate students who have learnt and have experience in teaching Operating Systems. RQ3 was evaluated in this phase: 1) measure the inter-reliability scores between raters, 2) measure the percentage score of raters that agree with the acceptability measure of the generated question, 3) evaluates the acceptability of the generated questions using the proposed question generation strategies in terms of grammatical error (RQ3.3), understandability (RQ3.4), relevancy (RQ3.5), usefulness (RQ3.6) and correct question type (RQ3.7). A questionnaire was used as a medium to get expert's feedback, and the quantitative statistical measure was used to calculate means and standard deviation for each of the question generation strategies of different question categories. The statistically significant differences for question generated between different QG strategies, and also between ontologies were evaluated in RQ3.8 and RQ3.9.

.

1.4.3 Phase 3

RQ4 evaluate the question selection module that uses topic-based indexing and randomized function for filtering question in AQGen. The purpose is to improve the feasibility of AQGen in term of question filtering. The feasibility is measured using accuracy, precision and recall.

1.5 Thesis Outlines

The thesis consists of seven chapters. The remainder of the thesis is constructed as follows.

Chapter 2 provides a review of the literature for automatic question generation, the use of ontologies as a source of keywords for generating questions and designing question templates for producing large numbers of questions. It begins by explaining about learning and assessment in tertiary education and continues with a discussion of the importance of questions as a mechanism to assess student's knowledge. This discussion is followed by with reviews on the current strategies for generating questions using ontologies as sources of keywords and question templates as a means to produce semantic questions. This leads to a further study of question generation strategies which maps the question template via ontology concepts.

Chapter 3 discusses the pre-analysis conducted to evaluate whether the ontology used in this work is appropriate for question generation purposes and the types of questions that could be generated using the ontology. There are two pre-analyses that are of interest in this work. First, to check either the concepts presented in ontology were the keywords usually used to assess students on Operating Systems. In order to validate the ontology, keywords in revision questions from an Operating System textbook were used. The analysis shows half of the keywords in that questions are similar to the concept names in the chosen ontology. The second analysis looked into types of questions that are generally asked for assessment in Operating Systems.

Chapter 4 discusses the process of creating question templates from the questions obtained in the analysis of question types in Chapter 3. There are 17 question templates from three question categories used in this work which have been finalised. Apart from question templates, this chapter also discusses the design of question generation strategies which use ontological approaches. There are four question generation strategies proposed using three ontological approaches.

Chapter 5 discusses the process of designing an Automatic Question Generator (AQGen), which is a tool developed to generate questions based on the proposed question generation strategies discussed in Chapter 4. AQGen was developed using the Java platform in two stages, with the first stage without filtering and the second stage with topic filtering. The first stage AQGen generated a lengthy number of questions, whereas topic filtering in the second stage allowed for topic-based question selection.

Chapter 6 evaluates the questions generated from AQGen proposed in Chapter 5. The tool was built and used to generate questions which were evaluated by subject matter experts in Operating System and Computer Networks & Security subjects. The evaluation examined two major aspects: i) the acceptability measure of the questions generation strategies, ii) the generalisation of the question generation strategies to be used on other ontology.

Chapter 7 discusses the updated version of AQGen, which introduce topic-based question selection. The topic-based indexing and randomised function for filtering question are discussed.

Chapter 8 is the final chapter and discusses the contribution of this research. The limitations and recommendations are also discussed, as well as the future work that could be undertaken following this research.

Chapter 2

Related Work

2.1 Learning and Assessment

Assessment is a process of gathering data, and it defines the way the instructors gather data about their teaching in response to their students' learning. An assessment has a significant impact on learning, as it provides data for further evaluation to establish benchmarks for education standards. The overall process of assessment involves gathering data, interpreting and making appropriate reports. According to Broadfoot and Black (2004), assessment is essential for several reasons, such as to provide learning feedback, allow the student to be graded or ranked, and to encourage learning. Apart from demonstrating a student's level of knowledge, the assessment provides an idea of the method that student will use to engage with the learned material (Biggs, 2002). Biggs has also distinguished the concept of 'assessment for learning', which provides feedback to improve the learning process, and 'assessment of learning' which is more suitable for reporting purposes. These concepts are similar to the formative and summative assessment. Formative assessment is used to monitor student progress such as question and answer sessions, tutorial, open quizzes and observation during in-class activities (Sadler, 1998; Yorke, 2003; Bennett, 2011). On the other hand, summative assessment is used to determine progression to the next level of a programme and is like a final examination (Dixson & Worrell, 2016). The outcome from the formative assessment is feedback from instructors (Nicol & Macfarlane-Dick, 2006), while summative assessment is usually produced grades.

Advances in technologies have served to shift traditional based assessment on paper to the electronic versions of the assessment. Early electronic assessment (e-assessment) question generation was implemented by populating questions from manually constructed assessment questions, which were kept in online question banks. Studies on this type of system have focussed on the algorithm of selecting related questions, which include using fuzzy algorithms (Sultan, 2010), utility-based-agent systems(Naz et al., 2010) and genetic algorithms (Xiong & Shi, 2010; Teo et al., 2012).

Assessment activities such as setting, responding and marking the assessment are expensive and consume significant amounts of time and effort for both students and instructors (Tosic & Cubric, 2009). According to a survey on teacher behaviour (Stiggins et al., 1986), teachers prefer informal methods more often than a formal approach, as informal methods require less preparation and skill. Formal methods include MCQ, T/F, short answer, and essay, while informal method includes such as tutoring questions and observation (Cunningham, 2005). This survey revealed that the process of constructing questions such as MCQ and short answer required more effort and time.

One essential way to operate assessment is by asking questions. The question is, in effect, a sentence that is used to request information. There are three types of sentence, namely simple, compound and complex sentences. A simple sentence has the most basic elements; a subject, a verb and key terms to be assessed.

Therefore, there is a need to automate the process of creating questions. This process requires knowledge of the specific domains and a set of rules to generate question automatically. Early work on question generation, uses Natural Language Processing (NLP) to process text into questions (Heilman & Smith, 2009; Singer & Donlan, 1982; Ali et al., 2010), keyword to questions (Zheng et al., 2011; Chali and Hasan, 2015), and query to questions (Lin, 2008). However, the drawback of these approaches is difficult in inferring implicit relations between keywords (Heilman, 2011; Hoshino & Nakagawa, 2005; Liu et al., 2005). The growth of Semantic web technologies has increased the application of automatic question generation. The capability of its knowledge representation allows for better knowledge reasoning process. Therefore, work in this thesis adopts automatic question generation using ontology. Some of the existing research that motivates this work is discussed in Section 2.5.

2.1.1 The educational value of questions

Questions are a powerful tool that provides different applications for different situations. Questions are widely used in everyday activities for people to find out something they did not know (Kearsley, 1976). Personalised user authentication in an online environment also uses questions as an authentication mechanism (Woo et al., 2016). In a learning environment, both learners and instructor may ask questions for different reasons. Instructors in this thesis context refer to teachers or lecturers that deliver teaching to students.

A question may help in knowledge construction(Chin & Brown, 2000) and help increase self-understanding(Chin, 2006; Graesser et al., 1992; Chi et al., 1994), by self-questioning the content covered in their lesson or tasks given by their instructors. Moreover, questions help instructors to validate students' conceptual understanding(Fensham et al., 2013) in order to provide appropriate feedback which is part of formative assessment (Bell & Cowie, 2001). Experiments examined by (Pedrosa de Jesus et al., 2005) in regards to university students during their group mini-project have revealed that active question asking increase interaction between students and instructors in the learning and teaching of chemistry. Another research study has found that active question asking will increase student motivation toward a topic that they learned(Chin & Kayalvizhi, 2005). To a certain extent, questions may provoke critical thinking for instructors and make them aware of the inadequate knowledge they have provided (Watts et al., 1997).

2.1.2 Types of questions

Questions are used for knowledge elicitation and to request some information. There are different types of question, depending on the types of assessment.

In this thesis, written assessment questions that are usually used for exercise questions, tests and examination questions are identified. Types of question that are usually used are Multiple Choice Questions (MCQ), True/False (T/F), Fill-in-the-blank (FIB), short/long answer questions (S/LAQ). Following are the description of the question types defined in (Cunningham, 2005).

Multiple Choice Questions (MCQ) are composed of one question with multiple choices of an answer, including the correct answer and distractors.

True / False questions are only composed of a statement which is responded to by indicating true or false.

Short answer questions are typically composed of a brief statement, and the written answer varies in length.

Long answer or essays questions provide complex statements that require long written responses.

In this thesis, S/LAQ are generated based on the question categories discuss in sections 2.2.1, and the reason for choosing S/LAQ over other question types is discussed in 2.5.1.

2.1.3 Model of question asking

PREG is a cognitive computational model of question asking developed by Graesser and his research group (Graesser & Olde, 2003; Otero & Graesser, 2001). PREG is developed for student question asking. Although PREG is aligned with the question generation that focuses on the instructor, however, PREG has two main components which are a set of production rules and a conceptual graph that could be adopted in this research. In this research, a set of rules were designed to map question templates with ontology concepts. Ontology is more expressive compared to the conceptual graph used in PREG.

2.2 Automatic Question Generation

Recently, an automatic question generation research has raised the interest of researchers from different disciplines, especially research in an educational setting. Developing automatic methods for Question Generation (QG) can alleviate the burden of both traditional and e-learning assessments. Research on question generation has been conducted for a long time (Piwek & Boyer, 2012) and one of the earliest works on question generation was proposed by Cohen (1929) that used an open formula with unbound variables to express the content of the question. According to (Rus et al., 2008), question generation is the task of automatically generating questions from various inputs such as raw text, database, or semantic representation. Unstructured text such as topic sentences, paragraph, and ontology are generally used as sources for QG.

The question generation process stated in Rus et al. (2008) involves four steps, which are: When to ask the question?, What the question is about? Question type identification, question construction. Question generation has been implemented for several reasons as mentions in (Le et al., 2014); 1) knowledge/skill acquisition, 2) knowledge assessment, and 3) provide tutorial dialogues systems. Thus, many applications from different domains have been deployed for several purposes, and some are discussed in section 2.2.2.

2.2.1 Question categories and question templates

There are four categories of tutorial question, as discussed in (Graesser & Person, 1994), which are the question with a short answer, the question with long answer, assertion, and request or directive question. The short answer has been further classified into several subcategories which are; verification, disjunctive, concept completion, feature specification, and quantification whereas long answer have classified further into the definition, example, comparison, interpretation, causal antecedent, causal consequence, goal orientation, instrumental/ procedural, enablement and expectation judgemental. According to the author, six of the question types; causal antecedent, causal consequences, goal orientation, enablement, expectation, and judgemental are highly correlated with the deeper level of cognition in Bloom's taxonomy. In order to capture different level and knowledge of learners, a question categorisation scheme is proposed by (Graesser et al., 1992) which categorized question-based on three dimensions: the content of information requested, the context of the question, and degree of specification. There are 17 question categories suggested and the list of categories, description based on (Erdogan, 2017) and examples are as follow:

1. Verification
    - Description: Question that requires yes or no response to factual questions.
    - Example: File type can be represented by file extension. True or false?
2. Disjunctive
    - Description: Question that requires a simple decision between two alternatives.
    - Example: Is the data-in or flow-in register that read by the host to get input?
3. Concept completion
    - Description: Question that requires details of the concept/term.
    - Example: Explain the term segmentation in memory management?
4. Feature specification
    - Description: Question that determines the qualitative attributes of an object or situation.
    - Example: What is the purpose of the deadlock avoidance algorithm?
5. Quantification
    - Description: Question that determines quantitative attributes of object or situation.
    - If there are 32 segments in memory, each of size 1Kb, then how many bits should the logical address have?
6. Definition
    - Description: Question to determines the definition of a concept.
    - Example: What is an operating system?
7. Comparison
    - Description: Question to identify similarities and differences between two or more object.

- Example: What is the difference between swapping and page relocation?

8. Interpretation
   - Description: A description of what can be inferred from a pattern of data.
   - Example: What is happening when thread cancellation occurs in multithreading?
9. Causal antecedent
   - Description: Asks for an explanation of what state or event causally led to the current state and why.
   - Example: What condition is required for a deadlock to be possible?
10. Causal consequence
    - Description: Asks for an explanation of the consequences for an event of state.
    - Example: What will happen if we remove the main memory from the system?
11. Goal orientation
    - Description: Question asking about motives or goals behind an action.
    - Example: Why segmentation is needed in memory management?
12. Instrument/procedural
    - Description: Question asking for plan or instrument to accomplish a goal.
    - Example: How do you convert an IP address to a logical address?
13. Enablement
    - Description: Asks for an explanation of the object, agent, or processes allow some action to be performed.
    - Example: Which system call creates a new process?
14. Expectational
    - Description: Asks about expectations or predictions (including violation of expectation).
    - Example: Why segmentation has not occurred?
15. Judgmental
    - Description: Asks about the value placed on an idea, advice, or plan.
    - Example: What do you think of the security in mobile apps?
16. Assertion
    - Description: Making a statement indication lack of knowledge or does not understand the idea.
    - Example: I don’t understand the security system
17. Request/directive
    - Description: Request for an action of some processes.

- Example: Would you convert IP address to logical address?

The web-based collaborative argumentation system proposed by Le et al. (2014) is a further example of a recent question generation application that uses Graesser et al. (1992) question classification. This work proposed a technical approach to assist students in understanding a given discussion topic by mean of generating questions for argumentation with the use of semantic information. The work employs a syntax-based, template-based and semantic-based approach where WordNet and Wikipedia are used to provide semantic information to generate a semantic question. Question templates defined according to Heilman (2011) were constructed for each of nine categories in Graesser and Person (1994). Each placeholder in the question template was replaced with a noun or noun phrase from a discussion topic. The question categories covered the overall topic for argumentation.

Another research study that adapts the question classification by Graesser and Person (1994) was for investigating the relationship between question types and how large the elicitation process is discussed in Bodoff and Raban (2016). This works also investigated how elicitation affect fee that is paid for mediates search services. They consider only four out of seventeen categories proposed by Graesser and Person (1994) due to the kind of questions encountered in their data sample. Apart from the four categories – *Concept Completion*, *Enablement*, *Instrumental* or *Procedural*, and *Quantification* – a new *Identity* question category has been introduced, which was found to be reliable in predicting elicitation and fee. Table 2.1 presents four question categories used by Bodoff and Raban (2016) in their works and the modification they take to fit their research purposes.

Table 2.1: Bodoff and Raban modification of question categories

No.	Question Category	Abstract Specification by Graesser	Bodoff and Raban Modification
1.	Short Answer		
	Concept Completion	Who? What? What is the referent of a noun argument slot?	This category is not only for short answers but includes long answers as well. There also include 'where' and 'when' is this category.
	Quantification	What is the value of a quantification variable? How many?	Include more elaborate quantitative analysis instead of a short answer question.
2	Long answer		
	Instrumental or procedural	What instrument or plan allows an agent to accomplish a goal?	Bodoff and Raban (2016) define this category as being like asking advice compare to Graesser that define it as a plan to accomplish the goal. Ex: *Are there any legal precedents which could be used in court to help support a mother's right?*
	Enablement	What object or resources allows an agent to perform an action?	Instead of asking *what object that will allow a person to act,* Bodoff and Raban (2016) twist it to *what object meets certain criteria*.

2.2.2 Question generation applications in a different domain

Automatic question generation may be classified into three methodologies which are syntax-based, and have a heavy reliance on Natural Language Processing (NLP), template-based models that use predefined question templates, and a semantic-based approach, which is usually domain dependent (Al-Yahya, 2011). Question generation tasks in educations are classified according to its purposes and are classified according to three classes: for knowledge/ skills acquisition, for knowledge assessment, and tutorial dialogues systems. Therefore, question generation has been implemented in various domains for different purposes, as follows:

A. Question generation application in the English language

Earliest works have a heavy reliance on Natural Language Processing (NLP) tools. Natural language parser has been used to analyse sentence syntax (Wolfe, 1976). In Brown et al. (2005), question generation was studied and implemented to generate questions for vocabulary assessment to measure student's vocabulary knowledge before they are assigned with appropriate text to read. The question generation works using the information available on WordNet (Fellbaum, 1998) and adapts Stahl's (1986) model of word knowledge generates six types of questions; definition, synonym, antonym, hypernym, hyponym, and cloze questions. The model explains three levels of knowledge, which are the association of vocabulary words with other new words, the comprehension of a vocabulary word in a particular context, and lastly the usage of a vocabulary word.

B. Question generation application in Computer Science

The Programming Language Online Exam Platform (PLOEP) was deployed to resolve the plagiarism issue amongst student in writing source code. According to Lin and Chou (2011), fixed exercises increase the chances of plagiarism, which reduces the learning effect. The generated ontology model, consisting of a concept part that used basic set theory and c++ implementation part has been tested and reported as being able to produce 3,624 open questions from four question templates. The high number of questions produced makes it possible to generate a dynamic question for each student and hence reduce the chances of plagiarism. However, this approach is expensive when implemented in other domains since it requires knowledge for connecting c++ implementation with other domains. It also requires an expert to design an appropriate question template that can be manipulated and reused using the knowledge presented in the ontology model.

C. Question generation application in Multimedia Systems

The paper presented by Castellanos-Nieves et al. (2011) discusses semantic annotation used for marking student answer for predefined test questions. Annotation for the question is provided by the teacher, together with the generated questions. An answer annotation is a semi-automatically generated ontological element which is detected from Natural Language Processing (NLP) using an incremental knowledge acquisition algorithm. There are three steps involved in this algorithm in order for the process to map linguistic expression with the knowledge entity namely preparation of a sentence until it gets tags using POS Tagging, search phrases that identify linguistic expression representing an ontological element and set in a context process that used grammar pattern to put the knowledge expression into context.

D. Question generation application in Biomedical

An unsupervised approach was investigated in Afzal and Pekar (2009), to generate MCQ questions using semantic relation from the biomedical text and classify them into a surface semantic pattern. The GENIA tagger in Tsuruoka (2006), mainly for the medical domain, was used to extract candidate patterns and Name Entity (NE) recognition. Three different pattern types were used for content words; untagged word pattern, POS-tagged word pattern and verb centred pattern. The results show that the POS-tagged word pattern and verb centred pattern have higher precision compared to untagged word pattern using Chi-square ranking method. The extension of their work was presented in Afzal and Mitkov (2014), who discusses the use of the unsupervised dependency-based approach to extract semantic relations to be applied in the context of automatic MCQ generation for the biomedical domain. The proposed techniques are used to identify the most important NEs and terminology in a text document and recognised semantic relation between them using the dependency tree model. Two domain experts evaluate the quality of the automatically generated MCSs based on readability, the usefulness of semantic relation, relevance, acceptability, and overall MCQ usability. Both experts asked for a score value to be given for the criteria above, and their agreement is measured using weighted kappa Cohen (Cohen, 1968). This system has employed sophisticated Natural Language Processing (NLP) at the beginning for information extraction from a text document to convert into the dependency tree of NE. Their evaluation results show that the MCQ generation system is capable of getting high precision rates.

E. Question generation application in Medical

The ESICT (Experience oriented Sharing of health knowledge via Information and Communication Technology) (Andersen et al., 2012) is a question answering system that provides information on diabetes mellitus using two sources: SNOMED CT a multilingual vocabulary of clinical terminology, diabetes medical documents. Question generation is part of this system features to generate potential users' questions regarding biomedical facts, diet, exercise, lifestyle, and diagnostics. In this system, sentences from related diabetes documents in the Danish language is extracted and transforms into interrogative questions using Tregex and Tsurgeon (Levy and Andrew, 2006), a tool for querying and manipulating tree data structures.

F. Question generation application across multiple domains

AutoTutor is an intelligent tutoring system inspired by human tutoring strategies (Graesser et al., 2012). AutoTutor is a project-based system developed by researchers affiliated with the Institutes of Intelligent Systems at the University of Memphis. This means that the AutoTutor is a combination of related systems that share specific software component and features of AutoTutor. The related system in AutoTutor has a related system that form a family of related system that cover various domains including conceptual physics (Graesser et al., 2003; Rus et al., 2013; VanLehn et al., 2007), computerized learning game (Halpern et al., 2012) , research ethics (Hu and Graesser, 2004), games for scientific enquiry (Millis et al., 2011), computer literacy (Graesser et al., 2004), and biology (Olney et al., 2012).

2.3 What is Ontology?

There are two different perspectives of ontology, from the viewpoint of Philosophy and the viewpoint of Artificial Intelligence (AI). `Ontology' is originally a technical word in philosophical jargon, which has now become trendy in AI. According to Gruber (1995) and Gruber (1993), ontology is a theory of what exists and is relevant in a specific domain, and of how it is organised. The notion of ontology that is presented in AI research focuses more on the conceptualisation or categorisation process, before the actual building of a knowledge base. The continuous growth of a web that integrates with business and personal life makes ontology research move to the web. The Paper on ‘Ontology Development 101’ (Noy & McGuinness, 2001) mentions several reasons for the need of the ontology be included in World-Wide Web. Ontology is a formal specification of a conceptualisation (Gruber, 1993) that can be used to share and enable the reuse of domain knowledge. The sharable and reusable ability of ontology makes it become widely used in a different domain (Richards, 2009; Brewster & O’Hara, 2007; Fernández-Breis & Martiinez-Bejar, 2002; Carchiolo et al., 2010).

Ontologies contain domain knowledge in the form of concepts, instances belonging to these concepts and relationship between them. In the educational setting, domain-specific ontologies or course ontologies were used as inputs for generating questions. Many ontology editors are available to support the construction of ontologies with several options of web ontology language for concepts representation.

2.3.1 Ontology Editors

Ontology editors are tools to support the construction of ontologies. There are many ontology editors available for web ontologies. Ontology editors can help to edit, import and visualise an ontology that is developed before and during the building of ontology. Moreover, ontology editors are useful for building ontology schemas (terminological component) alone or together with instance data. Protégé 4.1 is an integrated software tool used to develop Knowledge-based systems. Applications developed are used in problem-solving and decision making in a particular domain (Knublauch et al., 2004). This tool is an open-source application and has competency evaluation support at an instance level. It checks the relations that exist between concepts and instances. OilEd 3.5 is an ontology editor that helps the user to build ontologies using DAML + OIL language. It is available as an open-source project under the GPL license (Bechhofer et al., 2001). The user needs to have a reasoner in order to get a full benefit of OilEd. It comes with FaCT reasoner when we download and install this tool. However, this tool does not have competency evaluation support. OntoEdit version 6 is an ontology editor that supports the development and maintenance of ontologies using graphical means (Sure et al., 2002). The tool allows the user to edit a hierarchy of concepts or classes. This tool has a competency evaluation support that is available via an OntoKick plug-in. It does also allow for checks at the instance level, where it checks the relation that may exist between concept and instances.

2.3.2 Web ontology language

A web ontology language is a language for defining and instantiating web ontologies. There are several types of web ontology languages, such as XML, RDF, DAML, and OWL. XML Schema is a language for restricting the structure of XML documents. It provides a surface syntax for structured documents but imposes no semantic constraints on the meaning of this document. Resource Description Format (RDF) (Klyne & McBride, 2003) is a data model for objects and relations between them and provides a simple semantics for this data model, and these data models were represented in XML syntax. RDF Schema is a vocabulary for describing properties and classes of RDF resources, with semantics for generalisation hierarchies of such properties and classes. DAML is markup language for the semantic web (Hendler & McGuinness, 2000) The DAML program has generated the DAML+OIL markup language DAML+OIL is the syntax, layered on RDF and XML that can be used to describe sets of facts making up an ontology. DAML+OIL and its friend OIL (ontology integration language) use RDF namespaces to organise and assist with the integration of arbitrarily much different and incompatible ontology. OWL adds more vocabulary to describe properties and classes: among others, relations between classes, cardinality, equality, richer typing of properties, characteristics of properties and enumerated classes (Knublauch et al., 2004). In all, OWL has more facilities for expressing meaning and semantics than XML and RDF. It provides semantic data for agents on the internet.

2.3.3 Domain ontology for educational settings

Many ontologies have been constructed on different domains such as the FOAF ontology (Brickley & Miller, 2010) for describing data in social networking, the Gene Ontology (Ashburner et al., 2000; Consortium, 2016) for biomedical annotations, and Plant ontology (Consortium, 2002) which represent plants' concept, stages and anatomy. Ontology is developed for a different scope and purpose.

Many well-published ontologies are publicly available in the ontology library consortium; the ONKI ontology server (Viljanen et al., 2009) covers more areas such as business, culture, geography and thesauri, Protégé ontology library (Musen, 2015) for biomedicine, e-commerce, and organizational modelling, BioPortal ontology library (Noy et al., 2009) and the OBO Foundry (Smith et al., 2007) for biomedical ontology, and OntoSelect (Buitelaar et al., 2004) ontology library for supporting search mechanism. However, some ontologies are created for experimental purposes on particular research projects, and domain ontology has developed for educational purposes which cover subjects' syllabus is very limited. In this thesis, such domain ontologies for educational purposes are referred to as course ontology.

The ontologies in computer science that are publicly available are Operating System ontology and Computer Networks & Security ontology which are used in this work. The ontology of Operating System (Ma et al., 2006), is an ontology for a standard undergraduate operating system course. The ontology is built using OWL semantic language, and the concepts are extracted from four sources which include three textbooks (Silberschatz et al., 2002, Silberschatz et al., 2004, Tanenbaum, 2001) and lectures notes from one of the authors. It is reported that the ontology has more than one thousand concepts that spread into six parts; operating system overview, process management, storage management, I/O systems, Distributed Systems and Protection and Security.

Different operating system ontologies in Viljanen et al. (2009) contain 97 concepts and are spread into nine parts; concurrency control, file systems, fundamental of an operating system, I/O system, Linux system, memory management, process and thread management, protection and security, and system software. The ontology uses using OWL semantic language and the source for concepts selection is unknown. Therefore, in Chapter 3 of this thesis discussed the ontology concept appropriateness validation using keywords from OS assessment questions.

An ontology has also been built to explore knowledge in Computer Networks (Murugan et al., 2013). The ontology uses using OWL semantic language and consists of over 500 concepts that built with World Wide Web Consortium (W3C) standard (Berners-Lee et al., 2001). The ontology concepts have been organised into eleven parts; scope, scale, topology, communication media, OSI model, TCP/IP model, protocol, security, Network operating system, network hardware, and performance. It is reported that the ontology has about 550 relationships instances and 33 types of relationship. This ontology is used in this thesis, along with OS ontology by Viljanen et al. (2009) to show that the proposed question generation strategies could be generalised to other ontology.

2.4 Prior work on question generation using ontology

Advances in Semantic Web technologies have created interest among researchers in developing ontology-based applications in numerous research areas. One such research area is the field of assessment. The survey conducted has revealed a reasonable prospect of ontology related to e-learning (Al-Yahya et al., 2015). The survey conducted shows the current usage of the word ‘ontology’ search online between 2000 and 2012 increased significantly. Therefore, research in question generation using ontology as a semantic knowledge repository has gained interest among researchers.

2.4.1 The ontology-based question generator

The OntoWare system is developed by Holohan et al. (2005) and introduces an authoring environment for semi-automatic generation of the learning objects. It is deployed using Semantic Web technology, and thus the system may generate a question from ontology elements, which is the focuses of this research. This first version of OntoWare which focusing mainly on adaptivity and personalisation of a learning object was then extended to allow for the generation of assessment for problem-solving skills in the domain of relational databases (Holohan et al., 2006).

The OntoQue is an automatic question generator that generates assessment questions using domain ontology. The work that is discussed in this system is presented in Al-Yahya (2011) and implemented using a semantic-based approached. The system is capable of generating Multiple-Choice Question (MCQ), True/False (T/F) and Fill-In-the-Blank (FIB) questions using OWL domain ontologies to populate meaning assessment questions. Four domain ontologies were used for evaluation purposes; a self-created history ontology called *HistOnto*, Travel ontology for travel information and hotel (Knublauch et al., 2004), the *SemQ* ontology for Arabic vocabulary (Al-Yahya et al., 2010) and *People&Pets* (P&P) ontology for tutorial in OWL (Bechhofer et al., 2003) . Substantial research efforts have been made in the generation of ontology-based question generation. Three question strategies were proposed by Olney et al. (2012), which are; contextual verification, forced choices, and causal chain. Both contextual and forced-choice questions use features extracted from two related concepts for generating question stems. Olney also used three question types, which are *pumps*, *hints*, and *prompt* that bind with the low, medium and high level of specificity specified in (Graesser and Person, 1994).

2.4.2 Question generation approaches for Multiple Choice Questions

Ontology elements such as classes, instance and properties are exploited to automatically generate MCQ, T/F, FIB, S/LAQ. Based on the literature review (Papasalouros et al., 2008; Tosic & Cubric, 2009; Cubric & Tosic, 2011; Al-Yahya, 2014), the ontology-based question generation approaches may be classified into four main approaches:

- The individual concept-based approach: uses all individual assertions in the ontology.
- The terminology-based approach: uses the relation between the class and sub-class in the ontology.
- The class-based strategy uses the relationship between classes and individuals in the ontology.
- The property-based strategy: uses the relationships between instances/individuals in the ontology.

The first implementation of ontology-based MCQ question generation begins in 2008. The question generation proposed by Papasalouros et al. (2008) has introduced three approaches for ontology-based single response MCQ generation: class-based, property-based, and terminology-based. The proposed approaches generate questions according to the specified ontology represented in standard Web ontology language, OWL and hence can be used to generate questions for different ontologies as compared to template-based that might restrict the choice of a question asked. The work focuses on creating the correct (key answer) and incorrect statement (distracters) for the MCQ question that used the ‘Choose the correct sentence’ question stem. Approaches were constructed based on the ontology elements such as concepts, instance, properties and the relation between them. First, the class-based approach which generates distracters based on the concepts and their individual. There are five strategies defined under the class-based approach. Second, the terminology-based strategy which generates distracters based on the relationship between two concepts. There are two strategies defined under the terminology-based approach. Finally, is the property-based strategy, which uses a relationship between individuals in the ontology. There are four strategies defined under the property-based approach. The proposed approach has been evaluated using five different domain ontologies; *Eupalineio Tunnel, Msc Program, Travel v.1, Travel v.2,* and *Grid Resources*. The result of the experiment shown that class-based and terminology based provide better syntactically correct question but produce less amount of question compared to property-based questions. Much research on ontology-based question generation has been extended from their question generation strategies and Tosic and Cubric (2009) have optimised and implemented their strategies into Protégé ontology editor as a plug-in.

A new strategy for question generation to identify key answer and distracter for MCQ using text similarity measure and Semantic Web is presented in Cubric and Tosic (2011). The proposed strategies are a combination of annotations on ontology elements and question templates. MCQ ontology was developed as a basis for the development of the MCQ format specification. The question templates were extended from Holohan et al. (2005) question generation approaches, with the introduction of the new stem using annotated information. The question templates are as follows:

- Which one of the following definitions describe the concept A?
- Read the text x below and decide which one of the following concepts is a correct replacement for the blank space in the text.
- Read paragraph x below and decide which one of the following concepts it defines.

The key answer for MCQ was chosen from the highest similarity index, while distracters are chosen from the less similarity index. This approach was also implemented in the *OntoQue* system presented in Al-Yahya (2014) and Al-Yahya (2011).

Semantic-based approaches have been implemented in OntoQue system using domain ontology (Al-Yahya, 2011). The question generated based-on knowledge and recall level of Bloom's taxonomy (Anderson & Sosniak, 1994; Bloom, 1956). There are three question generation strategies deployed in the system which are class-membership based strategy, individual-based strategy and property-based strategies. The output of OntoQue was evaluated using precision which indicates the proportion of 'Good' question out of all generated items. High precision is obtained from the experiment for three ontologies except for *P&P* ontology due to the low numbers of class/individual ratio.

The annotation-based stems and strategies for generating MCQ distracters were introduced in Cubric and Tosic (2011), based on the assumption that the questions will become more difficult if the choice of distracter is similar. The system (SeMCQ) is developed as a plug-in for the Protégé ontology editor (Knublauch et al., 2004) uses domain ontology to generate MCQ item. The semantic similarity used the combination of text similarity measure with ontology elements similarity measure proposed in (Bach & Dieng-Kuntz, 2005). The 'question template' was used to mean the semantic interpretation that applies during the mapping process between domain ontology and the target MCQ ontology. The predefined 'question template' is designed to follow the level of difficulty in Bloom Taxonomy. The 'question template' was populated with various concept and annotation in the domain ontology. Predicate logic is used to represent a relationship of class and annotation in 'question template.' The restriction of this work is the relationship used in domain ontology and is restricted to inheritance or 'is-a' relationship and might not be able to represent specific course domain in detail. Moreover, the 'question template' is quite restricted and might need to be enriched to produce more choices of questions.

The work presented by (Alsubait et al., 2014) discusses the experience in generating ontology-based MCQ and proposed a difficulty-control theory that can reasonably provide good alternative answers by measuring the similarity between stem and distracter. Considerable different in varying the similarity between the correct and wrong answer for each question were used to set the difficulty of questions. The goal of the work presented was to evaluate how efficient the approach of varying the similarity between key and distracter was in controlling the difficulty of questions and hence how able the approach was to generate good questions. The experiment has been run on handcrafted ontology to generate questions candidate which gives 913 questions where 633 were an easy question, and the rest were difficult questions. *SubSim(.)* and *GrammarSim(.)* similarity measure proposed by (Alsubait et al., 2014) were used for specifying the difficulty of each question. Three domain experts for that particular course were asked to evaluate the 50 randomly selected questions, and the result shows that 46 useful questions were generated out of 50. Out the 50 questions, six were selected to be put on the test paper for a student to answer, and the system incorrectly categorised only one question's difficulty level. The work presented only generates question candidates, and a predefined question template has been used to demonstrate the result.

Unlike the previous researchers approaches that uses concepts and relationships between concepts in the ontology, a work proposed by (Vinu & Kumar, 2015) investigated difference strategies of generating MCQ questions where they utilised the terminology axioms in ontology such as universal, existential, and cardinality restriction on concepts. Furthermore, they have proposed two novel approaches i) node-level-set based approach ii) edge-label-set based approach to control the difficulty level of the generated MCQ. Vinu and Kumar (2015) have considered open-world assumption for distractors selection, where a systematic method used to choose distractors compared to other researches that randomly choose any instances from a class other than the correct answer as distractors. Similarity-based theory in (Alsubait, et al., 2013) is used in their research to measure the difficulty level of generated MCQ. A handcrafted ontology called Harry Porter Book was built by the research team and used as a running example.

The automatic question generation of MCQ type of questions was studied further in (Faizan & Lohmann, 2018) who exploit the use of the hierarchical relationship of semantic web dataset, which in this work used DBpedia. The idea proposed was similar to the terminology-based approach by Papasalourus et al., (2008) which select the distractor from the child class of the hierarchy. However, the work has been extended to generate different level of question difficulty based on the depth of hierarchical relations in the dataset. According to this work, the deeper level of the hierarchy, the closer those distractors related to each other. The simple example given in this work was triple information like ‘soccer-player’ is a subclass of ‘Athlete’, and ‘Athlete’ is a subclass of ‘Person’. Lionel Messi and Cristiano Ronaldo are the instances from ‘soccer-player’ would be chosen as a good pair of distractors, compared to Barack Obama (both belong to ‘person’ class).

Most recent work on MCQ generation can be found in (Leo et al., 2019). The source of knowledge The Elsevier Merged Medical Taxonomy (EMMeT) is a search-based application in a clinical setting with more than 900K concepts covering clinical areas such as drugs, anatomy, clinical findings, organisms, symptoms and procedures, 1.4M hierarchical and associative relations and over 30K semantic relations. Covering clinical areas like anatomy, clinical findings, drugs, organisms, procedure and symptoms This work suggests an automatic generation of medical case-based MCQ for medical education and medical licensing examinations. There are four question templates implemented in this work which are:

i) What is the most likely diagnosis?
   - There are three template’s stem entities for this question: *Patient Demographics, History* and *Symptom*, and the option entities are *Diseases*. Two object properties *have a risk factor,* and *hasClinicalFinding* are also used to relates between entities.

ii) What is the drug of choice?
   - There are two template’s stem entities for this question: *Patient Demographic* and *Symptom*, and the option entities are *Drugs*.

iii) What is the most likely clinical finding?

- There are two template's stem entities for this question: *Patient Demographic* and *Symptom*, and the option entities are *Clinical Finding*.

iv) What is the differential diagnosis?

- There are three template's stem entities for this question: *Patient Demographics, History* and *Symptom*, and the option entities are *Diseases*. Three object properties *have RiskFactor, hasClinicalFinding* and *hasDifferentialDiagnosis* are used to relates between entities.

The results reported that their approaches could generate the four medical case-based MCQ successfully and highlighted different area the ontology can be enriched.

There is a lack of higher-order thinking types of questions generated from most of these works. Their approaches only test students' ability to recall learned information. This is because the question templates used, are of simple sentences such as 'Choose the correct answer" and "What is X?". The answer option provided make it the recall level of questions. The research by except for the work proposed by Leo et al. (2019), proposed MCQ for higher-order thinking but the question template is only suitable for medical ontology as it captures knowledge such as symptoms, disease and clinical finding. Therefore, in this work, S/LAQ is implemented using the ontological question generation approaches. In section 2.5.2, there will be further discussion on how the S/LAQ being implemented using ontological question generation approach.

2.5 The value of the ontology-based approach to Short/Long Answer Questions (S/LAQ)

Assessment questions are comprising of many types. All the types of questions have still been used for assessment questions. The selection of the types of questions depends typically on the knowledge of learner that needs to be tested, how easy it is to construct, how easy it is to mark, or the type of question that usually is used by the previous instructor to test students. Each question types have advantages in assessing different learning outcome. The promotion of web technology for e-learning has shown the emerging used of MCQ, T/F compared to the other types of questions in the e-learning system. This is due to the capability of the automatic marking using e-Learning system compared to other types of question such as short/long answer which requires lengthy words for answers. However, many instructors still prefer to create S/LAQ for their assessment due to numbers of reasons. In subsection 2.5.1, we discussed four reasons as to why S/LAQ is worth to consider for automatic question generation. In the following subsection 2.5.2, we explained how this work would differ from the previous work presented in section 2.4.2 and finally in subsection 2.5.3, and we presented how the learners and teachers could benefit from this work.

2.5.1 Reasons for autogenerating S/LAQ

There are four main reasons for autogenerating S/LAQ compared to other types of questions.

- Retrieving more accurate information concerning student learning

Assessment question is a tool to analyse student performance and different questions type assessing different knowledge level. Unlike MCQ and T/F question, there is no guessing on the answer in S/LAQ, and the student must supply the answer to the questions (Funk & Dickson, 2011). Thus, the information regarding student learning is more accurate with more specific detail on the problematic part that students are facing.

- Motivate student to adopt more effective study techniques

Students were motivated to study differently and more effectively when they prepare for short answer examination (Balch, 2007). The outcome of this study is consistent with the result of the similar study by (Pinckard, et al., 2009) on the effect of examination question format toward student performance. The result of the study shown that a more challenging learning environment and student motivation toward effective learning techniques could be created by utilising short answer questions.

- Demonstrate student's creativity

Both short and long answer questions provide an opportunity for the student to demonstrate their creativity by throwing their ideas based on their understanding. This also means that scoring can be quite subjective, and students could gain marks even for any related answer, unlike MCQ that choose the most suitable answer. This indirectly influenced student to learn how to write a good answer to this kind of questions that enhanced student creativity in argumentation of idea or concept.

- Test difference level of cognitive

Short answer questions are more structured questions that often used to test basic knowledge and key fact about a subject. Short answer questions would generally be a fill-in-the-blank or definition questions which fall under low level of cognitive such as recall and recognition. Whereas long answer question is less structure, therefore are favoured to test the higher level of cognition such as analysis, synthesis and evaluation.

Therefore, the auto-generation of S/LAQ is proposed to give a variety of type of questions that could be asked in the e-learning system. The generation of the answer is not included in this work, but the question generation strategies proposed, would help an instructor to generate as many questions as possible to enable dynamic question generation. Dynamic questions would allow a different set of question to be executed at a short time and could affect student learning performance. This strategy also will replace the way of creating a question in the e-learning system, where questions previously were manually inserted into the assessment system or randomly selected from the question database.

2.5.2 How the taken approaches differ from the previous work

A. Question types and question templates

In this work, the question strategies are applied to S/LAQ. In this work, 17 question templates have been suggested. It is derived from textbook review questions as well as online past year exam assessment questions in Operating System subject, and the 17 question templates are shown in Table 4.1. The question templates cover three question categories which are *definition*, *comparison* and *concept completion*. The question templates include a question with one keyword and two keywords. Questions with two keywords will require relation between a concept in ontology to control the question scope. As compared to the Papasalourus et al.,(2008) works, he proposed a question generation approach for finding correct answers and distractor for MCQ type of question. One MCQ question templates suggested which is “Choose the correct sentence”. Therefore, the question templates for S/LAQ are different from the MCQ question templates.

B. Domain ontologies

Unlike other research on question generation approach that used a handcrafted ontology, this work has implemented using publicly available ontology from ontology library on the web. By implementing this approach, the suitability of existing ontology for question generation could be validated. Two validation of ontology are conducted i) the completeness of ontology concepts compared to the concept in the Operating System subject, ii) are the representation of the ontology elements adequate for question generation purposes.

C. Strategies of mapping question template to ontology elements

Question generation strategies proposed by Papasalourus et al. (2008) used hierarchical relation between a concept in ontology to generate answer options for MCQ. The answer option is taken from the hierarchical relation with a format `[subconcept] [relation][superconcept]`. The following example shows how keywords from ontology are selected for MCQ answer options.

Given the snapshot of the hierarchical ontology relationship as follow:

```
mouse is an input-device

keyboard is an output-device

scanner is an output-device

printer is an output-device
```

This representation shows two concepts; (i) input device and (ii) output device, which both have sub-concept (mouse) and (keyboard, scanner and printer) respectively. The correct answer option can be selected from the first line of representation (mouse is an input-device), while the distractors were chosen from sub-concept from concept (ii). Therefore, the MCQ generated is as follow:

Choose the correct sentence.

a. Mouse is an input device (correct answer)
b. Keyboard is an input device (distractor)
c. Scanner is an input device (distractor)
d. Printer is an input device (distractor)

In contrast, the question generation strategies proposed in this work used relation in ontology to capture the relationship between two keywords. The relationship could be in the form of hierarchical relation or sibling relation. The relationship of keywords is very important in order to generate a useful question. A question that has two keywords that is not related may generate an inappropriate question. Consider the following example of comparison type question:

i) Differentiate between computer and noodle.
ii) How do Facebook and the University of Warwick differ?

The first comparison questions contain two keywords to assess which are “computer” and “noodle”. This question is structurally correct, but it is not appropriate and meaningless as the two keywords are not related and of a different domain; one is a technology device, and another one is food. The second question also tries to compare “Facebook” which is a technology platform and “The University of Warwick” which is an educational department. There is no main idea to differentiate between two keywords in those two questions.

Good comparison question compared two pieces of related keywords. The relatedness between keywords can be in the form of in the same methods, in the same components or same ideas. For example:

iii) Differentiate between the input device and output device.

The third question can be deemed as syntactically and semantically correct question. This question is comparing two types of devices in a computer system. The two keywords “input device” and “output device” have a relation as a computer device. Thus, capturing the relationship of keywords is an essential process in developing an automatic question generation system. This piece of information can be captured using ontology representation, and the following triples can be executed and process to represent the relationship of keyword with another keyword.

```
Input-device is-a computer-device

Output-device is-a computer-device
```

The triples indicate that both the input device and output device are related by sibling relationship with computer device as a superclass concept. Comparing both two keywords would avoid generating meaningless comparison questions. It keeps the boundary or scope of the comparison, which at the same time reduce question ambiguity.

Next, consider the concept completion types of question which use hierarchical relation in the ontology.

iv) Explain the term segmentation in memory management.

Segmentation and memory management are the two concepts in the question. The preposition "in" in the question indicates that the keyword "segmentation" in the "memory management" topic. In order for this question to be meaningful, the keyword "segmentation" must be fall under "memory management" in a hierarchical relation. When the relationship between keywords is not observed, and any keyword can replace "memory management", the question will be meaningless. For example, if "keyboard" is used instead of "memory management", the question becomes "Explain the term segmentation in keyboard". The generated question is inappropriate and meaningless.

Therefore, in this work, experiments were conducted to investigate the suitability of the ontological approach (selection of concept as well as hierarchical relation) to capture relatedness of keywords in a question and using the question templates to create an acceptable question. The question template gives the semantic meaning of the overall question. The question template like "differentiate" and "explain" provides the purpose and objective of the assessment questions.

The contribution of this work lies in the extension of automatic question generation using ontological approaches for S/LAQ. The S/LAQ is still widely used in assessing student knowledge and to help students expanding their knowledge thru question inference. Furthermore, as the advancement in educational technology promotes online assessment, including S/LAQ, autogenerated questions will provide a different way of assessing students. The approach suggested in this work will promote the automatic generation of dynamic questions.

2.5.3 How this work would benefit teachers and learners

S/LAQ is widely used in assessing student learning. Most of the examination in university specifically computer science courses still using short and long answer question types in their assessments. There are also a vast amount of practice or revision questions of a different subject that can be found on the web. Although some researchers claim that S/LAQ is easy to constructs, the amount of time that instructors could take to generate a set of assessment questions could be used for giving student feedback. Some courses could have a few quizzes, tests and weekly tutorial per semester. Manually constructing questions might expose to delay due to instructor commitments, mood, and sickness. With this automatic question, generator instructor can generate questions anytime with a few clicks. Therefore, autogenerating S/LAQ will give an impact to the instructor in term of effort and time. This work also can be a guideline for generating more complex question structure in future.

Learners might benefit from AQGen as a revision tool. According to the literature in (Graesser & Person, 1994), one way of effective revision is by a question asking. By regularly responding to questions would increase learners' cognitive level. With this automatic question generator, a question from all topics will be covered in a system without having to search for different material. Most of online and textbook have limited revision questions. With this question generator, learners should expect access to thousands of revision questions.

2.6 Summary

This chapter has explained the theories and components relating to the assessment question and automating the process of constructing them in the educational setting. It has provided an overview of traditional assessment questions, which are manually constructed by the instructor to the e-learning system that populated manual questions for automatic selection by the system and later automating the process of question construction. The main purpose is to alleviate the instructors' burden to construct question manually and at the same type to introduce the generation of a large number of questions to enable the dynamic selection of questions. Questions are an important tool to assessed students' knowledge and are required in all levels of education. There is a requirement to create various types of question to cater for different levels of knowledge. Various research studies have been conducted to explore the automatic generation of questions using different types of question and from different sources and for different purposes.

Section 2.1 has discussed the components for generating questions and the question asking model that builds on our framework. Section 2.2 explained the question generation process and its application in different domains. This section also discusses question categories and question templates. Graesser et al., (1992) question categories are adapted in this research due to the categories of question found in the subject evaluated. Section 2.3 discussed the overview of the ontology as an input source for knowledge in term of the standard, language and ontology editors. The domain ontology was also explained through the example of ontology used in this thesis evaluation. Finally, section 2.4 has examined the approached of question generation in previous works.

The preliminary analysis of the literature indicates that most of the question generation strategies were explored for finding distracters of MCQ questions, and some are to extract concept for T/F and FIB questions. Most of the works investigate how to improve the similarity between concepts to increase questions difficulties. However, to the best of our knowledge, no research explores the automatic question generation explicitly for short and long answer type of questions for formative and summative assessment in education. Question generation in this thesis is implemented to assess student's acquisition of factual knowledge using concepts represented in the ontology.

The next chapter will present a further exploration of the use of OS ontology for question generation. The analysis of the literature review will be used for discussion and the construction of a question generation framework. Therefore, we present some important terminologies that will be used in the following chapters of this thesis.

The question type is different formatting for asking questions. The examples are MCQ, T/F, S/LAQ and FIB and the detailed discussed in section 2.1.2. In this work, we explore the automatic generation of S/LAQ.

Question template is a sentence with a placeholder. It is used in this work as a question structure which the placeholder filled with appropriate keywords from the ontology concept. The question template is from the OS revision questions which have been extracted and formalised as discussed in section 4.2. There are 17 finalised question templates, as presented in Table 4.1.

Question category is a classification of the question according to the purpose of the question asked. There are several classifications and the most reviewed and used is the classification suggested by Graesser et al. (1992), as discussed in section 2.2.1. However, based on the analysis we get from the survey on existing OS review questions, as discussed in section 3.4, only five categories from the classification identified. In this works, we choose three categories which are *definition, concept completion* and *comparison*. The 17 question templates are designed according to these 3 question categories, and we called it categorised question templates. The list is presented in Table 4.1.

Question generation strategies is a way of mapping ontology concepts with question templates. We proposed four ontology-based question generation strategies, as discussed in section 4.3. Each categorised question templates obtained will fall under one of this question generation strategies. Finally, the full combination list of the question generation strategies and the categorised question templates is presented in Table 4.1.

Chapter 3

Pre-Analysis: How well Ontology can be used for Generating Assessment Question?

3.1 Introduction

Ontologies have been used in many research projects for generating assessment questions, as discussed in Chapter 2. This kind of research aims to benefit instructors by providing support and intelligent assistants for the automatic generation of questions. However, existing ontologies are not designed mainly for this purpose, and the concern is that an ontology will not be competent enough to act as a semantic source for the question generation process. A pre-survey has been undertaken to identify whether there is a possibility that questions can be generated from an ontology for automatic question generation.

The research question that will be answered in this chapter is:

RQ1: How well can an ontology be used for generating factual assessment questions?

In order to answer RQ1, experiments have been conducted to evaluate the appropriateness of the chosen ontology for question generation. This is done through an information retrieval technique to measure string similarity between keywords that exist in questions with the concept's name in the ontology. Besides, existing available assessment questions on the Internet were used to identify question categories that are usually used for assessing student knowledge and which of those that could be generated using an ontology. The following are the sub research questions to evaluate how well the chosen ontology can be used to generate questions automatically.

RQ1.1: Do the course ontologies evaluated in this research contain concepts that are appropriate for the subject being evaluated?

RQ1.2: What are the categories of question that can be generated using the ontology?

3.2 Preparing data input

There are two data inputs used in this pre-evaluation, namely assessment questions and an ontology. Assessment questions are used in the analysis as a competency question to identify whether or not the ontology contains concepts that are usually used as keywords in assessment questions. It is also used to identify which question categories were used in the subject being evaluated, which in this case is the Operating System subject in the Computer Science domain.

3.2.1 Assessment questions

There are two sources of assessment questions used in this work, which are revision questions from an Operating System textbook (Silberschatz et al., 2012) and past year final exam questions of Operating Systems.

Revision questions consist of a total of 263 short answer and true/false questions across 15 chapters in the textbook. The questions range from simple to complex and vary in level of knowledge. The categories of question include verification, comparison, definition, examples and concept completion questions.

Past year final examination questions were taken from the Universiti Teknologi MARA question bank. There are nine sets of question from the years 2010 to 2015. The sets of questions were labelled based on the subject code, which is CSC520 and followed by the year of the examination.

3.2.2 Course Ontologies

Course ontologies may be categorised as domain ontologies where the scope is limited to delivering educational learning content. There are some course ontologies in the Computer Science domain found in the literature as discussed in section 2.4.3, but the competency of the ontology to be used as a source of semantic information for automatic question generation is not known. Therefore, we consider validation of course ontologies which are used as a source of keywords to be assessed for automatic question generation.

An Operating System subject from the Computer Science domain was chosen to begin with, due to the availability of this course ontology on the web and the nature of the test questions being mainly factual. An Operating System ontology from the (Viljanen et al.) was chosen to be evaluated and has been named as ontoOS throughout this thesis. OntoOS have 97 concepts with hierarchical relations and contain nine main concepts, as shown in Table 3.1.

Table 3.1: Main concepts and their number of subclasses in ontoOS

No.	Concepts	No. of Sub Concepts
1	Concurrency control	11
2	File systems	7
3	Fundamental of operating system	6
4	I/O_ systems	19
5	Linux system	7
6	Memory management	11
7	Process and thread management	6
8	Protection and security	15
9	System software	6

3.3 Do the Ontologies Contain Concepts that are Appropriate for the Subject Being Evaluated?

In order to determine whether the concepts in ontoOS are appropriate for generating a question for Operating System subject, an experiment has been conducted to identify the existence of Operating System subject keywords inside ontoOS. A technique called competency questions was used to check the coverage of the ontoOS concepts. The use of competency questions is a well-known technique for determining the requirements the ontology should fulfil. Competency questions used by (Grüninger and Fox, 1995) evaluate whether an ontology is sufficient for its intended purposes. Therefore, in this work, competency questions were used as a means to check whether the concepts represented in ontoOS were appropriate for generating questions the OS subject. Competency questions used in this experiment are the OS review questions in the (Silberschatz et al., 2012) textbook. Review questions were used as competency questions to ensure that concepts in ontoOS are the keywords that generally used for assessment in Operating Systems.

3.3.1 Pre-processing of question sentences using n-gram model

Additional steps were taken before conducting similarity checks to allow matching between n-arrays of strings in questions. This is because most of the keywords that exist in questions are in the form of single words (such as *interrupt*), a pair of words (such as *Operating System*) and some are more than two words (such as *twisted pair cable*). Question likes "What is an Operating System?" and the matching concepts in the ontology, "Operating-System", need further pre-processing in order to extract the matching keyword and concept. The n-gram model is used for pre-processing before calculating the similarity coefficient. An n-gram, in this case, refers to a sequence of tokens or words. For example, the question "What is an Operating System?" is a 5-gram. In order to calculate the similarity between the keywords in a question and the concepts in an ontology, the questions were divided into tokens using the n-gram model, where n-grams used in this experiment were trigram, bigram and unigram. Use of these trigrams, bigrams and unigrams were intended to cater for different forms of keywords like single words, pairs of words and triple words respectively. Examples of the trigram, bigram and unigram structure of the following question are as follows:

"What is an Operating System?"

Trigram :

{* * What, * What is, What is an, is an Operating, an Operating System?}

Bigram:

{* What, What is, is an, an Operating, Operating System?}

Unigram:

{What, is, an, Operating, System?}

Symbol '*' is added to the bigram and trigram, so the total number of token for each array cell is 2 and 3 respectively.

After dividing the questions into several tokens, the similarity scores were calculated for each of the elements inside this three n-gram against each concept inside the ontology. The similarity scores were calculated using Sørensens Similarity Index, and the similarity score adapted for this experiment was 0.90. In the next section, we will discuss the calculation for the similarity score in detail.

3.3.2 Calculating similarity coefficients using Sørensens Similarity Index

Sørensens Similarity Index, also known as the Dice's Similarity Coefficient (DSC) algorithm, is applied by extracting character bigrams to calculate similarity scores, s, of two strings. The similarity coefficient for two strings, K and C, which represent a keyword in a question and concept in the ontology, respectively, is calculated as follows:

$$s = \frac{2n_t}{2n_K + 2n_C} \qquad (3.1)$$

where n_t is the number of character bigrams found in both strings, n_K is the number of bigrams in string K and n_C is the number of bigrams in string C.

The similarity score between the keyword in question, K, and the concept in ontology, C, varies between 0 and 1. A similarity score of 1 indicates both K and C intersecting or matching and 0 if there is no intersection at all. For example, to calculate the similarity score between “WIRED” and “WIRELESS”, the set of character bigrams in each word can be broken down as follows:

{WI, IR, RE, ED}

{WI, IR, RE, EL, LE, ES, SS}

The first set has four elements, the second set has seven elements, and the intersection of these two sets has only three elements: WI, IR, RE. Inserting these numbers into the formula, the calculation of similarity score is as follows:

$$s = (2 \times 3)|(4 + 4) = 0.55$$

Therefore, the similarity score between “wired” and “wireless” is 0.55, which indicates the two strings are not equal.

On the other hand, consider the question “What is an Operating System?”. The term ‘Operating System’ appearing in a question can be matched with the ‘Operating-System’ concept in the ontology. Pre-processing was performed to create a combination of words from each sentence and the first word for each question has an empty string as a pair which is denoted as an asterisk. For example the question “What is an Operating System?” is tokenised into a pair-wise string as [* What, What is, is an, an Operating, Operating System?] and stored in an array. The following algorithm is run to obtain the similarity score.

```
score = 0.0
for every question Q do
   split question Q into n-gram
   while has more tokens
      for every concept C in ontology
         calculate s
```

```
          if( s > 0.9)
             display concept C and its score
          endif
       endfor
    endloop
 endfor
```

Table 3.2 shows part of the similarity score obtained from matching the question's bigram and the concept "Operating System" in the ontology using formula 3.1.

Table 3.2: Sample of Similarity score obtained

Bigram	Concept	Score (s)
*** What**	Operating System	0.00
What is	Operating System	0.00
is an	Operating System	0.00
an Operating	Operating System	0.62
Operating System?	Operating System	0.97

Therefore, the last element of the question's bigram is matched with the "Operating System" concept in the ontology with the similarity score above 0.90. The similarity score of 0.90 was chosen to allow for the one or two different letters, such as question mark in this case.

3.3.3 Results for similarity score between keywords in questions and the matching concepts in the ontology

There are many ways for naming concepts in ontoOS, and the numbers of each occurrence are calculated using the similarity score. Table 3.3 shows the numbers of keywords of every single word, pair-wise, and triple terms obtained from the similarity coefficient calculation with the similarity score greater than 0.90.

Table 3.3: Categories of concept naming and their occurrences.

No.	Categories	*Num.*
1	Single terms	9
2	Pair-wise terms	47
3	Triple terms	1

The results show that out of 97 concepts in ontoOS, more than half of the concepts were used in the review questions. This indicates that the ontoOS has good coverage of the concepts for generating questions about Operating Systems. The rest of the concepts were not matched for several reasons such as the keyword not being asked in the questions and concepts represented in the form n-grams with n > 3.

Since most of the concepts existed in the review questions, the concepts presented in ontoOS can be used to generate assessment questions in the Operating System subject. After analysing the suitability of concepts in ontoOS, the next step was to identify which review questions could be generated using the ontology by formalising the questions into categorised question templates. This process may help in identifying and designing suitable question templates for assessment questions generated from the ontology.

3.4 What are the Categories of Question that can be Generated using an Ontology?

This work adopts a question template approach for automatic question generation. In order to identify assessment questions that can be generated from an ontology, the 263 Operating System (OS) reviewed questions (Silberschatz et al., 2012) were analysed. These questions have various question categories which assess different levels of knowledge.

3.4.1 Question's Structures

A question is a sentence used to request information. In English, there are three types of a sentence which are simple, compound and complex. A simple sentence has the most basic elements; a subject, a verb and a complete thought.

An assessment question **Q** is made up of 3 tuples (QW, P, KW), where QW is a question word, P is a Predicate, and KE is a Key element.

A question word is a keyword used to indicate the type of information needed, the predicate is the verb to express the action of the subject, and the subject is usually a noun in a sentence.

For example:

> ***Q***: What is an operating system?
>
> ***QW****: What*
>
> ***P:*** *is an*
>
> ***KW****: Operating System*
>
> ***Q***: Explain the advantage of Shortest Job First algorithm."
>
> ***QW****: Explain*
>
> ***P****: the advantage of*
>
> ***KW****: Shortest Job First algorithm*
>
> ***Q***: In the context of memory management, what is an address binding?
>
> ***QW****: In the context of…. what*
>
> ***P****: is an*
>
> ***KW****: memory management, address binding.*

The questions address factual and descriptive types of knowledge. They can be range from the definition, to comparison, to concept completion. The questions that ask for an opinion, common sense, general knowledge and inter-domain question will not be covered in this work.

3.4.2 Procedure for question classification

All the questions were classified into question taxonomies as proposed in Graesser and Person (1994). There are four main question taxonomies which are: questions with short answers, questions with long answers, assertions and request or directive questions. Short answer question is further classified into several subcategories, which are: verification, disjunctive, concept completion, feature specification and quantification, whereas long answer questions are classified into a definition, example, comparison, interpretation, causal antecedent, causal consequence, goal orientation, instrumental/ procedural, enablement and expectation judgemental.

3.4.3 Results for question classification

Table 3.4 shows the result of classifying the review questions into Graesser question categories. The work focuses on the question types that are used in the review questions for the evaluation in the later stages of this work.

Table 3.4: Classification of review questions into Graesser question category.

QUESTION TAXONOMY	NUM. OF QUESTION
Definition	7
Comparison	4
Concept Completion	110
Verification	61
Feature Specification	40
Others	41

The analysis has shown a few question categories used in the review questions. Concept completion was the most asked in the review questions. The structure of the questions used for student learning ranged from simple to complex, and the type from factoid to descriptive. There has been work on ontology-based question generation for verification types of questions, however, to the best of our knowledge, there is still no ontology-based question generation using templates that focus on short and long answers for definition, comparison, concept completion and feature specification.

However, not all questions in each category may be generated from an ontology due to the complexity and the expressiveness of the questions' sentences. Therefore, certain restrictions on questions need to be imposed to enable the questions to be generated from an ontology.

3.4.4 Results for the question that could be generated from ontology

The results in this section show the questions that could be generated using an ontology with the question templates. Questions were manually selected based on the type of questions described in section 3.4.1. The sample questions have been selected from various sources which include OS review questions and past year exam questions, as shown in Table 3.5.

Table 3.5: Sample questions

No.	Definition:	Source
Q1	What is an interrupt?	OS Review Question
Q2	What is a Process Control Block (PCB)?	OS Review Question
Q3	What is a controller?	OS Review Question
Q4	Define a circular wait condition in deadlock.	CSC520/2011
Q5	Define resource abstraction.	CSC520/2013
Q6	What is a safe state?	CSC520/2010
Q7	Define the following terminology: i. File ii. File Manager iii. File System	CSC520/2012
	Comparison	
Q8	What is the difference between protection and security?	OS Review Question
Q9	Differentiate between classic processes and modern processes.	CSC520/2013
Q10	Differentiate between mechanism and policy in CPU Scheduling.	CSC520/2014
Q11	Compare between deadlock prevention and deadlock avoidance.	CSC520/2014
Q12	Distinguish between a process and a thread.	CSC520/2012
	Concept Completion	
Q13	Describe a state diagram.	CSC520/2010
Q14	Explain what virtual memory is.	CSC520/2010
Q15	What is the function of the kernel?	CSC520/2010

3.4.5 Inappropriate representation of a concept's name

The analysis has identified the inappropriate representation of concepts that affect the generation of good questions. The inappropriate representation problem is divided into four different subcategories, and following discusses each of the categories with examples, why it causes problems and possible solutions.

A. Long concept name

Example:

interaction-between-processes-and-o-s is-a process-and-thread-management

Why it causes problems:

This kind of naming will give too much information for a concept. The long name representation will make it difficult to do reasoning. It creates problems in generating questions as this representation includes nouns as well as action verbs in one word. For example, if we generate a definition question, it may generate:

"*Define interaction-between-processes-and-o-s in process-and-thread-management".*

Recommended representation:

A simple name for concepts may be appropriate and avoid adding a verb to the noun to represent a concept.

B. A concept name contains two or more nouns.

Example:

- *O-s-services-and-components*

Why it causes problems:

This kind of naming will introduce ambiguity, as we are not sure of which concept is the key element for the question. For example, if we generate a definition question, it may generate: "*Define o-s-services-and-components*."

Recommended representation:

Use single concepts by splitting the compound concept into two.

C. Repeating the word used for sub and super concept.

Example:

- Security problem is-a security

Why it causes problems:

Although it is not wrong to represent concepts in this way, without pre-processing the concept name before generating the question, it will generate redundant words in the question. For example:

"Explain security problems in security."

It would be better to rephrased as:

"Explain problems in security."

Recommended representation:

As the super-class already mentioned about 'security', the suggested representation is *problem is-a security***.**

D. The naming of the concept includes an action verb.

Example:

Implementation-of-access-matrix

Why it causes problems:

This kind of word is not suitable for naming concepts. Usually, a concept will use a noun or noun phrase for representation. When we try to instantiate a definition type of question template that has a pattern "Define [X]." will create a question like:

"*Define implementation-of-access-matrix."*

However, if we represent the concept using a noun (access-matrix) and action verb (implementation-of) as object properties, then we get questions like "*Define access-matrix*." and "*Explain the implementation-of access-matrix*." respectively.

Recommended representation:

The action verb is more suitable to be used as an object property instead of combining with a concept word.

3.5 Designing Question Generation Framework

In this work, we have identified three challenges in generating an automatic question from an ontology for educational purposes. The task is divided into three stages, and the conceptual framework is shown in Figure 3.2.

3.5.1 Stage 1: Pattern Acquisition

The objectives are to generate categorised question templates and strategies to map question templates with ontology elements. First, the assessment questions obtained from revision questions as well as universities' question banks were analysed and validated using ontoOS to identify whether the concepts in the ontology represent keywords that normally used to assess the student in Operating System subject. The evaluation did not check for how complete the ontology can be used for question generation; instead, it is used to investigate whether there is a high possibility of the question being generated using ontology through appropriate concepts represented in the ontoOS. After investigating the possibility of question generation using the ontology, the next task was to formalise the questions using predicate logic to obtain a pattern for categorising the questions.

At the same time, we identified how the keywords in questions are represented in the ontology, especially in terms of how one keyword relates to another in the ontology. Identifying the relations is useful in deciding the strategy for question generation based on the template that we have. Although the keywords can be extracted from the ontology concepts, not all question templates are suitable for each of the concepts in the ontology, especially for a question that contains two or more keywords. For example, the question ‘Explain overlay in swapping' might not make sense since overlay only happens in virtual memory. Therefore, a strategy for generating a question that contains two related concepts is needed and proposed.

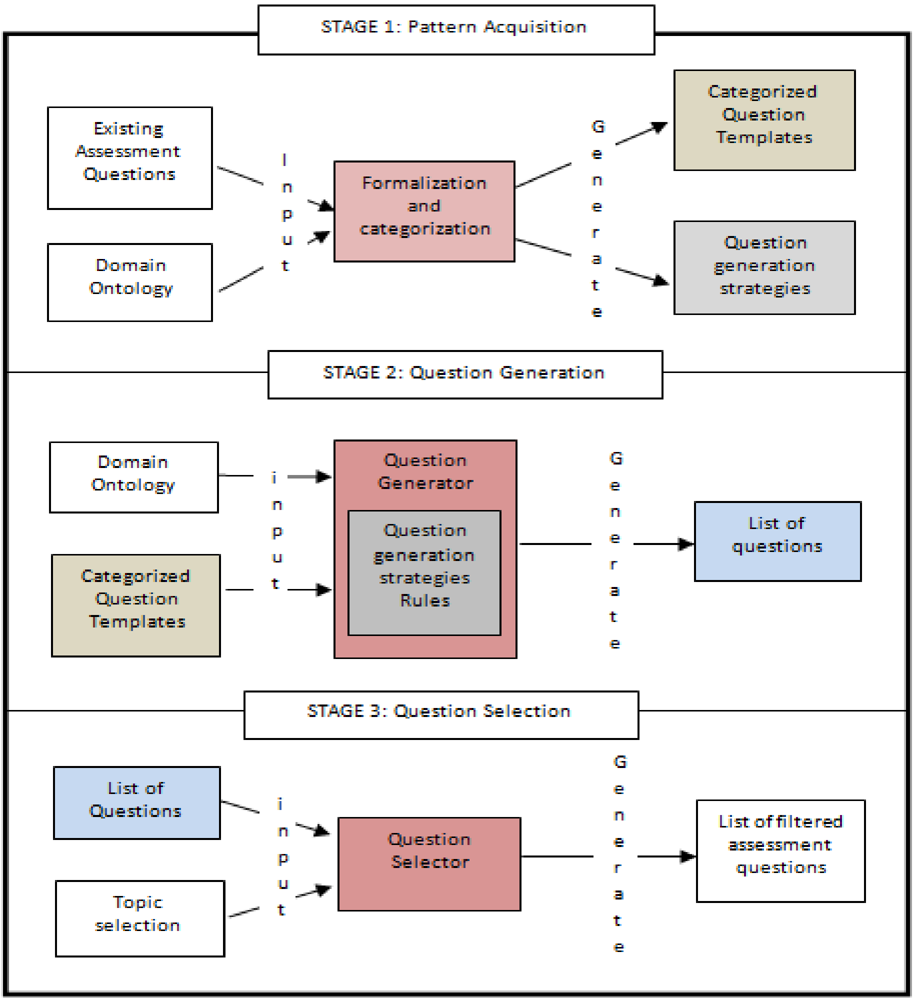


Figure 3.1: Question generation framework

The hierarchical ontology structure, which relates two concepts, may be used as a strategy to control the selection of concepts to create this kind of question. Besides, there might be a question where two keywords that are related in such a way that both need to be on the same topic in order to distinguish between them. For example, the question "Differentiate between segmentation and swapping" needs 'segmentation' and 'swapping' to be under the same topic of 'memory management' in order for the question to be correct and useful. Asking a question like 'Differentiate between segmentation and deadlock' will not be appropriate, although it seems correct. Therefore, exploring the strategies that are suitable for mapping question templates with ontology concepts will be another challenge to explore in this research.

3.5.2 Stage 2: Question Generation

In stage 2, the objective is to develop a tool that can generate questions based on question templates and strategies obtained from stage 1. In this stage, we start with designing the conceptual model of the tool, which includes the knowledge base structure as well as the rules applied for matching question templates with the ontology concepts. There are two modules in this system, which are Data Reader for reading the data file, and Question Generators for generating question using the three ontological approaches as follows:

- TermQG contain rules for generating terminology-based questions,
- ClassQG contain rules for generating individual-class-based questions, and
- PropertyQG contains rules for generating property-based questions.

3.5.3 Stage 3: Question Selection

Finally, assuming all the templates and strategies work well, the long list of assessment questions will be generated from an ontology. This will create a problem known as over-generation. To overcome this problem, we will explore how to select a suitable set of questions out of the list in our third stage of this research.

In this stage, we further enhance the features of the question generator to allow the instructor to enter the assessment required to generate a set of assessment questions. Here, a lengthy list of generated question in stage 2 will be filtered out, and the selection of questions will be based on instructor choice of topics. The selection approach will reduce the over-generation problem of questions where only the required questions will be created for each instructor. At this stage, we also resolve any shortcoming from the previous stage.

Chapter 4

Designing Question Templates and Question Generation Strategies

4.1 Introduction

This chapter discusses the generation of question templates by semantically interpreting existing assessment questions from the textbook and university past years' exam questions. We also identify suitable ontological strategies that match identified question templates. At this point, there are four question generation strategies have been identified using three question categories for question templates, these being *definition*, *concept completion* and *comparison*.

The research questions for investigation in this chapter are:

RQ2: How can questions be generated from course ontology?

RQ2.1: How should categorised question templates be constructed?

RQ2.2: What ontological question generation strategies can be applied to generate questions?

The objectives are to generate categorised question templates and strategies to map question templates with ontology elements. After investigating the possibility of question generation using an ontology, we proceed to formalise the questions using predicate logic to obtain a pattern for categorising the questions.

4.2 Designing Question Templates using Predicate Logic

After investigating the possibility of question generation using an ontology, the next step is to formalise the questions using predicate logic. This section discusses how to design a question template by using an ontology as a source of knowledge. There are five steps involved in designing question templates from the revised questions from the Operating Systems book and past year questions, as presented in Table 3.5. A simple formalisation of a sentence using predicate logic is applied to obtain the question templates. The next subsections discuss the steps needed to generate these question templates.

4.2.1 Collecting and selecting a question

The Operating System reviewed questions were collected from the textbook (Silberschatz et al., 2012). There are 256 questions of various question types and depth of knowledge. In this thesis, the interest is only to investigate descriptive and comparison types of questions. Thus only those types of questions were selected. The questions selected were of the following three criteria.

i. Simple questions that contain no more than three keywords,
ii. Descriptive (such as definition and concept completion) and comparison type of questions.
iii. Asking factual information.

4.2.2 Question Classification

Based on the results obtained in Section 3.4.4, three categories of questions were asked in assessments and could be generated using an ontology for operating systems. Therefore, selected questions were classified into three categories, where two are of descriptive type, and one is comparison type, according to the Olney et al., (2012) and Graesser and Person (1994) question classification.

The classification is based on the question words used in the questions. The question words which are used are listed in Table 3.5, which shows the questions that can be generated using the ontology and can be summarised as follows:

Definition type of question

- What is a/an
- Define

Comparison type of question

- Differentiate
- Distinguish
- Compare
- What is the difference

Concept completion

- Explain
- Describe
- What

4.2.3 Removing questions with the same pattern

Questions with the same pattern were removed to allow unique question templates to be stored in the question template repository. "Same pattern" in this context refers to a set of questions that all tokens in the sentence match except for the keyword, as shown in the following examples.

Example 1:

What is the Operating System?	*Selected*
What is Security?	*Removed*
What is Scheduling?	*Removed*

Example 2:

What is the difference between non-preemptive **and** pre-emptive scheduling algorithm?	*Selected*
What is the difference between protection **and** security?	*Removed*
What is the difference between an unsafe **and** safe state?	*Removed*

In example 1, the three definition questions have the same pattern, which is '**What is'** and differ in their keywords which are Operating System; Security; Scheduling. In example 2, the comparison questions have a pattern '**What is the difference between … and …?**' and substitute with different keywords. Therefore, for each of the examples, one question will be selected, and the other two will be removed.

4.2.4 Grammar and question structure checking

All questions in Table 3.5 had undergone grammar and question structure validation manually by a native speaker to ensure valid and correct question templates generated at the end of the process.

4.2.5 Removing question key terms and question words

Here, the question words and the key terms of question sentences are replaced with placeholders labelled [QW] and [C] respectively. Question key terms are then replaced with a variable to form question templates. If more than one key term exists in the questions, [C1], [C2] and [C3] are used to indicate different key terms.

For example:

Question: What is the Operating System?
Question Template: What is C?

Question: What is the difference between non-preemptive and pre-emptive scheduling algorithm?
Question template: What is the different between C1 and C2?

The next step is to remove the question word to create formalised question templates. For example, given the following question templates:

What is C?

What is meant by a C?

The formalised template becomes:

QW C?

Example 2: With question templates:

What is the different between C1 **and** C2?

Describe the differences between C1 **and** C2?

Compare C1 **and** C2?

The formalised template becomes:

QW C1 **and** C2?

However, the three question categories have different question structures. Instantiation of QW is general and would not be suitable for all question categories. For example, the QW in the formalised template ‘QW C?’ cannot be instantiated to use the comparison type of question since the structure of the question only uses one concept, C. Comparison cannot happen with only one concept being used. Therefore, categorised templates are more suitable for question generation with different question categories. The formalised question templates for the three question categories are as follows:

Definition:

- QWD C?
- QWD C.

Comparison:

- QWC C1 and C2?
- QWC C1 and C2.

Concept Completion:

- QWCC P C?

4.3 Ontology-based Question Generation Strategies

An ontology is represented as a set of concepts or classes in a subject area and shows their properties and relations between them. The relations of concepts in an ontology are mainly organised in a taxonomy which organises concepts into sub-super concept tree structures. Few research studies have proposed question generation strategies using knowledge in the ontology about domain entities such as classes, properties, individuals and relations between classes and individuals for generating MCQ, T/F, and FIB questions (Al-Yahya, 2014). Another work on question generation strategies used the relation between two ontology elements to create a key and distractors for MCQ types of questions (Papasalouros et al., 2008). The relation between different ontology elements has been categorised as class-based, term-based, and property-based. In this work, short or long answers to questions would also require a relation between two ontology elements.

Identifying knowledge about ontology concepts and relations is useful in deciding the strategies for question generation using categorised templates obtained from the previous section. Therefore, this work exploits the knowledge representation of the ontology elements to design question generation strategies for definition, concept completion, and comparison types of question.

Question generation strategies in this work are defined as rules to choose a concept or a pair of concepts from ontologies to be key terms for assessment questions. Question generation approach proposed by Papasalouros et al. (2008) was adopted in this work. The difference was on the strategies of mapping question template to ontology elements, the questions' types and the question templates used.

The process of identifying question generation strategies begins by analysing how keywords in a question are presented in ontologies. Assessment questions Q1 to Q15 from Table 3.4 were analysed, and the following subsections discuss the question generation strategies for choosing concepts from ontologies to be those key terms for the assessment questions.

4.3.1 Question Generation Strategies 1 (QGS1)

By using this strategy, the question can be generated *if there is a concept **C** which exists in an ontology.*

This strategy is used for assessment questions that only require one key term to be assessed. Examples of these questions are:

- What is an interrupt?
- What is a Process Control Block (PCB)?
- What is a controller?
- Define resource abstraction.
- Describe a state diagram.
- Explain what virtual memory is.

This strategy allows any concept that exists in the ontology to be a key term for the assessment questions. It could generate as many questions as the number of concepts in the ontology with a single question template.

The number of questions that can be generated is calculated as follows:

$$Total\ Questions = NumOfCon \times NumOfQT \qquad (4.1)$$

where

NumOfCon: Number of object property relation in the ontology

NumOfQT: Number of Question template

Consider the OS ontology, which has 97 concepts, thus 97 questions with the same question template but different in the key terms that could be generated. The CNS ontology consists of 537 concepts. This strategy is suitable to be used with the definition and concept completion type of question. Figure 4.1 and Figure 4.2 presents snapshots of the OS Ontology and CNS ontology respectively. Some assessment questions can be generated from both ontologies.

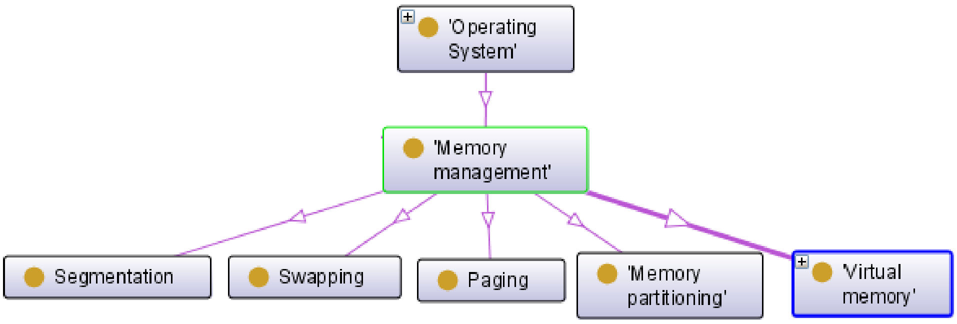


Figure 4.1: Snapshot of OS ontology representation

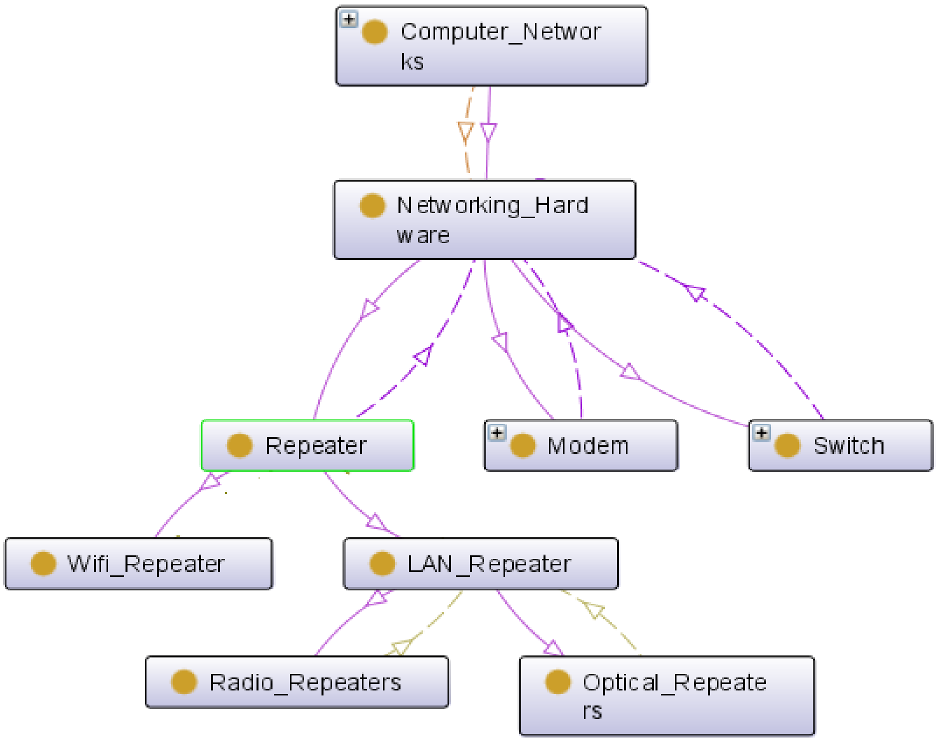


Figure 4.2: Snapshot of CNS ontology representation

The definition type of question that can be generated from the OS ontology includes:

- Define segmentation.
- Define memory management.

The definition type of question that can be generated from the CNS ontology includes:

- Define switch.
- Define modem.

The concept completion type of question that can be generated from the OS ontology includes:

- Explain what segmentation is.
- Explain what memory management is.

The concept completion type of question that can be generated from the CNS ontology includes:

- Explain what switch is.
- Explain what modem is.

4.3.2 Question Generation Strategy 2 (QGS2)

By using this strategy, the question can be generated *if concept **C1** is a subclass of another concept **C2** exists in an ontology.*

This strategy is used for assessment questions that require one key term to be assessed, but with extra information as to which main key terms, it belongs to. Examples of these questions are:

- Define a circular wait condition in deadlock.
- Explain client/server in the Network Operating System.

This strategy allows any hierarchical relationship that exists in the ontology to be a key term for the assessment questions. It could generate as many questions as the number of hierarchical relations in the ontology with a single question template.

The number of question that can be generated using this strategy is as follows:

$$Total\ Questions\ =\ NumOfHRel\ \times\ NumOfQT \tag{4.2}$$

where

$NumOfHRel$: Number of hierarchical relationships in the ontology

NumOfQT: Number of Question template

Consider the OS ontology that has 86 hierarchical relations, thus 86 questions with the same question template, but different in key terms could be generated, while the CNS ontology could generate 433 questions with its 433 hierarchical relations. This strategy is suitable to be used with the definition and concept completion type. By using the OS ontology as in Figure 4.1, the following assessment questions can be generated.

The definition type of question that can be generated includes:

- Define segmentation in memory management.
- What is paging in memory management?

The concept completion type of question that can be generated includes:

- Explain segmentation in memory management.
- Explain paging in memory management.

For these questions, ‘memory management’ is the main key term used to provide extra information to the key term being assessed (segmentation and paging). In order to generate these types of question, a strategy that looks into hierarchical relations is used. In this example, the key terms ‘segmentation’ and ‘paging’ must be the sub-concepts of ‘memory management’. While using the CNS ontology as in Figure 4.2, the following assessment questions could be generated.

The definition type of question that can be generated includes:

- Define modem in networking hardware.

- What is a switch in networking hardware?

The concept completion type of question that can be generated includes:

- Explain repeater in networking hardware.
- Explain switch in networking hardware.

4.3.3 Question Generation Strategy 3 (QGS3)

By using this strategy, the question can be generated *if there is a concept **C1** and concept **C2** that are subclasses of concept C exists in an ontology*.

This strategy is used for assessment questions that also need to assess two key terms. Examples of these questions are:

- What is the difference between protection and security?
- Differentiate between classic processes and modern processes.
- Differentiate between mechanism and policy in CPU Scheduling.
- Compare between deadlock prevention and deadlock avoidance.
- Distinguish between a process and a thread.

This strategy allows any two sub-concepts that belong to the same concept in the ontology to be used as a key term for assessment questions. This strategy is suitable to be used with a comparison type of question. This type of question needs to have two keywords that could go in the same category. For example, football and basketball which are both in the sports category.

The number of questions that can be generated using this strategy is as follows:

$$Total\ Questions = NumOfSRel \times NumOfQT \tag{4.3}$$

where

NumOfSRel: Number of sibling relationships in the ontology

NumOfQT: Number of Question template

The OS ontology has 198 sibling relations. Thus 198 questions with the same question template but different key terms could be generated. The CNS ontology contains 1188 sibling relations.

The comparison types of question that can be generated from the OS ontology as in Figure 4.1 include:

- Differentiate between switch and repeater.
- Distinguish between radio repeaters and optical repeater.
- Concerning LAN repeater, explain the difference between radio repeaters and optical repeaters.

The comparison types of question that can be generated from the CNS ontology as in Figure 4.3 include:

- Differentiate between segmentation and paging.
- Distinguish between swapping and paging.
- With regard to memory management, explain the difference between swapping and paging.

4.3.4 Question Generation Strategy 4 (QGS4)

By using strategy 4, the question can be generated *if there is a concept **C** that has object property **P** existing in an ontology.*

This strategy is used for assessment questions that require one key term to be assessed but using the object property relation. An example of this question is:

- What is the function of the kernel?

This strategy uses the information of the object property and its concepts in the ontology to create a question. In the example above, 'kernel' is the concept connected using the object property relation 'is the function of'. This strategy could generate concept completion types of question.

The number of questions that can be generated using this strategy is as follows:

$$Total\ Questions\ =\ NumOfPRel\ \times\ NumOfQT \qquad (4.4)$$

where

NumOfPRel: Number of object property relation in the ontology

NumOfQT: Number of Question template

There is no object property relation in the OS ontology. The CNS ontology has 38 object properties that makeup 562 triples, thus could generate 562 questions with the same question template but different key terms. Adding another one new question template could generate another 562 questions. This strategy is suitable to be used with the concept completion type of question. The following is the snapshot of the CNS Ontology, and some assessment questions can be generated. Referring to Figure 4.2 and Figure 4.3, the following assessment questions can be generated.

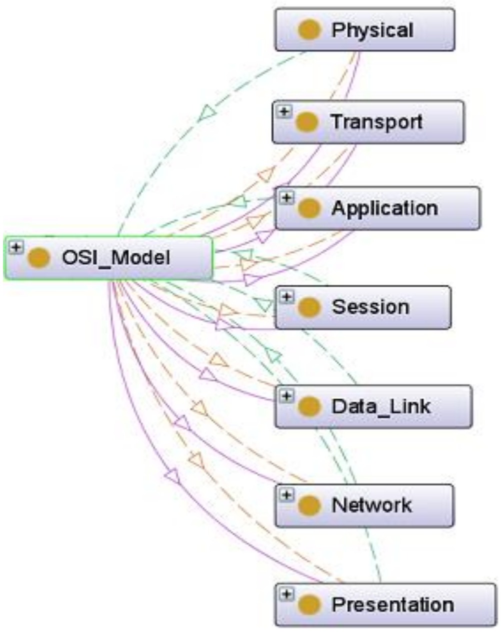


Figure 4.3: Snapshot of CNS ontology representation for OSI Model

The concept completion type of question that can be generated includes:

- What are the types of LAN Repeater? (from Figure 4.2)

- What are the layers of OSI models? (from Figure 4.3)

In the example above, both of the questions use the object properties 'types of' and 'layers-of' to generate assessment questions.

4.4 Proposed question generation strategies and their corresponding question templates

In this research, we proposed question generation strategies for different question templates. The strategy depends on the number of key terms used in a question, and if a question requires two key terms, a relation between key terms is investigated. In Section 4.2, we proposed categorising question templates to allow a variety of question words to be used to ask questions in each question category. In Section 4.3, we proposed question generation strategies using ontology concepts and relations between ontology concepts. A question generation strategy is a rule used to map question template with the ontology elements. Table 4.1 shows the combination of question generation strategies and the question templates that can use to generate assessment questions.

Table 4.1: Question generation strategies and question templates. [C], [C1], [C2] represent a concept in ontology, [SC] is super-class of the concept and [P] is the object property.

Strategy	Category	Template
QGS1	Definition	Define [C].
		What is [C]?
		What does it mean by [C]?
	Concept Completion	Explain the term [C].
		Discuss what [C] is?
	Comparison	*[no template]*
QGS2	Definition	What is a [C] in [SC]?
		Define [C] in [SC].
		In the context of [SC], what is [C]?
		What is meant by the term [C] in [SC]?
	Concept Completion	Within the context of [SC], explain the term [C]?
	Comparison	*[no template]*
QGS3	Definition	*[no template]*
	Concept Completion	*[no template]*
	Comparison	Differentiate between [C1] and [C2].
		What is the difference between [C1] and [C2]?
		How does [C1] differ from [C2]?

		Describe the differences between [C1] and [C2].
		Compare between [C1] and [C2].
		Concerning [SC], explain the difference between [C1] and [C2].
QGS4	Definition	*[no template]*
	Concept Completion	What [P] [C]?
	Comparison	*[no template]*

There are 16 question templates for OS subjects, 17 question templates for CNS subjects, and 4 question generation strategies proposed for evaluation in the next phase of the research. Two ontologies were used as a data input for the key terms, which are OS ontology and CNS ontology. Table 4.2 shows the number of questions that can be generated from the proposed combinations of question generation strategies and question templates for 3 question categories.

Table 4.2: Number of question that can be generated using strategies and question templates shown in Table 4.1

Strategies	No. of Question Template	Ontology	No of Question Generated
QGS1	5	OS	485
		CNS	537 x 5 = 2685
QGS2	5	OS	430
		CNS	433 x 5 = 2165
QGS3	6	OS	1188
		CNS	73 x 6 = 438
QGS4	1	OS	none
		CNS	562

The idea of proposing a question generation strategy and categorised question templates is to generate a massive choice of questions. By using 17 question templates, both OS ontology and CNS ontology could generate 2665 and 5850 questions respectively, including the inappropriate representation of concepts discussed in section 3.4.5.

4.5 Summary

In this chapter, we have proposed 17 question templates from 3 question categories. We also proposed question generation strategies which are the rules for mapping question templates to ontology concepts. In the next chapter, we discussed the development of a question generator to allow for the implementation process of the proposed strategies for evaluation purposes.

Chapter 5

Automatic Question Generator (AQGen)

5.1 Introduction

In this stage, we develop a question generator that will take an ontology as a source of keywords for the questions. This generator will generate all possible questions using question templates for each concept in the ontology. The proposed ontological strategies are applied in this question generator. The deliverable from this stage is a question generator that can generate a list of questions for three question categories, namely definition, concept completion and comparison.

5.2 Methodology

We start by designing the conceptual model of the prototype, as presented in Figure 5.1.

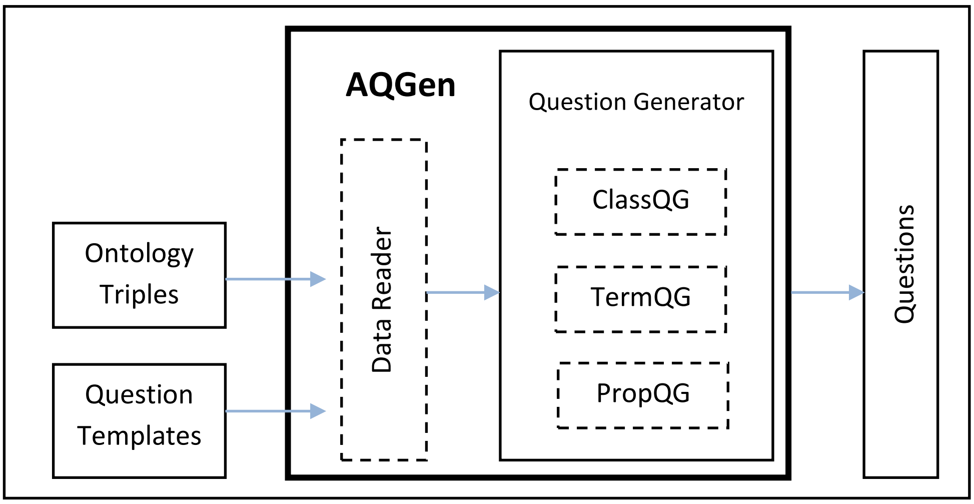


Figure 5.1: AQGen conceptual model

A tool AQGen has been developed using JAVA to implement the conceptual model and create the engine for the question generator. AQGen consists of 3 main components which are data input, reader and question generator. The data input structure and the question generation strategies rules for mapping question templates with the ontology elements are discus in this chapter. The following section will discuss the components of AQGen in detail.

5.2.1 Data Input

There are two types of data input that are used for this experiment, which are two ontologies and 18 question templates. Both data inputs are stored in the form of text files.

A. Ontology triples

A course ontology is a subject domain ontology that represents knowledge of educational learning content and, like other ontologies, it contains the concepts and relationships that exist between these concepts. An Operating System (OS) and a Computer Network & Security (CN) course from the Computer Science domain were chosen due to the availability of the course ontologies on the web and the nature of the test questions for this subject which are mainly factual. The OS ontology from (Viljanen et al., 2009) and the CN ontology from (Murugan et al.,2013) was chosen to auto-generate questions for this evaluation and are named as *ontoOS* and *ontoCN* respectively throughout this thesis.

The ontology elements are first extracted and represented as triples. Ontorion Fluent Editor 2015 software is used to get the structure of the triple for AQGen data input. For ontologies, triples are extracted in the form of "`concept1 relation concept2`". There are 97 triples in ontoOS and 562triples in ontoCN.

B. Question templates

There are 17 question templates that have been finalised from the three question generation approaches and three question categories. The question templates are constructed from a question word (such as ‘Define’, ‘What’) and concept [C1], [C2] and [SC]. The list of question templates is presented in Table 5.1.

Table 5.1: Question Templates

Approach	Category	Template	filename
Individual class-based	Definition	Define [C1]. What is [C1]? What does it mean by [C1]?	QTS_Def_Class.txt
	Concept Completion	Explain the term [C1]. Discuss what [C1] is.	QTS_CComp_Class.txt
	Comparison	*(no template)*	
Terminology-based	Definition	What is a [C1] in [SC]? Define a [C1] in [SC] ? In the context of [SC], what is [C1]? What is meant by the term [C1] in a [SC]?	QTS_Def_Term.txt
	Concept Completion	Within the context of [SC], explain the term [C1].	QTS_CComp_Term.txt
	Comparison	Differentiate between [C1] and [C2]. What is the difference between [C1] and [C2]? How does [C1] differ from [C2]? Describe the differences between [C1] and [C2]. Compare between [C1] and [C2]. With regards to [SC], explain the difference between [C1] and [C2].	QTS_Com_Term.txt
Property-based	Definition	*(no template)*	
	Concept Completion	What [P] [X]?	QTS_CConp_Prop.txt
	Comparison	*(no template)*	

5.2.2 Data Reader

This component is used to read an input file and pre-process the content before it can be processed further in the generator modules. In order to enable this process, some pre-processing tasks need to be carried out. This process is mainly used to process data input so it can be used in the question generator module later. There are three steps involved in pre-processing the data inputs, which are extraction of data inputs, tokenising triples and question templates, and assigning an index to the variables.

A. Extraction of data inputs

   There are two data inputs used in AQGen which are ontologies (namely ontoCN and ontoOS) and question templates, as shown in Table 5.1. Ontology triples are in the format of `[C1] [R] [C2]`, where `C1` and `C2` are concepts that have relationship R. Both data inputs were extracted and stored in arrays `ontoArray[]` and `queArray[]`.

B. Tokenisation

   The tokenisation process is performed on ontology triples in `ontoArray[]` to split and create an array of unique concepts and relations, in order to create a matrix table of relations in the next step. These are named `uniqueArray[]` and `relationArray[]` respectively and are used to keep indexed information of concepts and relations. Both `C1` and `C2` may be combined to form unique concepts and remove duplicates.

C. Creating a matrix table to store triple information

   A two-dimensional matrix table is created to store triple information to be used in the generator modules in the next step. The ontology file is read, and the relationship is stored in `relMatrix[indexC1][indexC2] = indexR`. `[indexC1]` and `[indexC2]` are the location index for `C1` and `C2` which stored in `uniqueArray[]`, while `indexR` is the location index for relation [R] stored in `relationArray[]`.

5.2.3 Question Generator

The generator component is further divided into three sub-modules which are individual class-based, terminology-based and property-based. Each of the sub-modules generates questions based on the four question generation strategies that have been proposed. The main task of this module is to instantiate variables in question template with the concept in an ontology. For example, ‘What is [X]?’ as question template, and generator module will instantiate [X] with concepts from the ontology using individual class-based sub-module.

The array location index of the variable is read and recorded before instantiation. The location for the first token of the question is set to 0. The location index for [X] needs to be recorded and later a concept from the ontology is assigned to that location index. In this case, the location index for [X] is 2 and a new concept from the ontology will be assigned to location index 2 in the question template.

Let `m` be the location index of variables: `conceptC1=”X”`, `conceptC2=”Y”`, `conceptSC=“SC” and propP=”P”`, the following algorithm is executed to store index information. [SC] is a variable used to indicate super class and is used in the individual-class-based module.

```
for m=0 to m=queArray[] size
    if(queArray(m) = = conceptC1)
          then assign m to idx_C1
    if(queArray(m) = = conceptC2)
          then assign m to idx_C2
    if(queArray(m) = = conceptSC)
          then assign m to idx_SC
    if(queArray(m) = = propP)
          then assign m to idx_P
```

A detail explanation of the mapping of question templates with concepts in the ontology using individual class, terminology-based and property-based approaches is presented.

ClassQG is the individual class module that is used to generate QGS1 questions. This approach uses a simple rule where the keyword from the question is replaced with an ontology concept. Question categories that can be generated using this approach are definition and concept completion. Examples of question templates for both question categories are as follows:

Definition: Define [X].

Concept Completion: Discuss what [X] is.

There is only one concept variable used in this example which is [X]. Therefore, the location indices for the definition question are 1 and concept completion is 3. After the location index has been identified, each of the unique concepts in the ontology is assigned to the identified location index. This module will generate questions such as "*Define Operating System*" and "*Explain what Operating System is*." Each question template from this module will generate as many questions as the number of concepts in the ontology. For example, ontoOS has 97 concepts; therefore, 97 questions could be generated by each question template.

TermQG is the terminology-based module that is used to generate QGS2 and QGS3 questions. Both QGS2 and QGS3 generate question using the hierarchy and the sibling relation in ontology, respectively. A hierarchy relation generates question using two concepts that have an 'is-a', 'is-an', 'has-a' or 'has-an' relation, while a cross-relation generates question using two concepts under the same parent in the concept hierarchy. The terminology-based approach using hierarchy relation can generate definition and concept completion questions, and sibling relations are used to generate comparison questions. Examples of question templates are as follows:

Definition: What is [X] in [SC]?

Concept Completion: In the context of [SC], explain the term [X].

Comparison: Differentiate between [X] and [Y].

A hierarchy relation in an ontology is indicated by the 'is-a' and 'is-an' relationship, and the index is stored in relationArray[] as 0 and 1 respectively.

Therefore, the hierarchical relation between [X] and [SC] is determined using relMatrix[X][SC]. The algorithm for instantiating a question using the QGS2 is as follows:

```
if relMatrix[X][SC] is equal or less than 1
      then queArray.set(X, uniqueArray(idx_X)
            queArray.set(SC, uniqueArray(idx_SC)
```

A sibling relation in an ontology is indicated by an ‘is-a’, and ‘is-an’ relationship with two triples and the index is stored in relationArray[] as 0 and 1 respectively. Therefore, the sibling relation between [X] and [Y] is determined using relMatrix[X][SC] and relMatrix[Y][SC], where both [X] and [Y] is a subclass of [SC]. The algorithm for instantiating a question using the QGS3 is as follows:

First, calculate the number of subclasses for each concept. Let m be the index of the super-class and n be the index of the subclass. The following algorithm extracted all triples of sibling relation and stored it inside `subArray[]`.

```
for m = 0 and m < size of relMatrix
  for n = 0 and n < size of relMatrix
      if relMatrix[m][n] is equal or less than 1
         store all subclass n in subArray[].
```

If the number of subclasses in subArray[], is two or higher, then perform the instantiation process by replacing variables [X] and [Y].

```
If subArray[] size is equal to or greater than 2
   queArray.set(SC, uniqueArray(m));

for a = 0 and a is less than subArray size
   for b = a+1 and a is less than subArray size
       queArray.set(X, con.get(a));
       queArray.set(Y, con.get(b));
```

PropQG is the property-based module that is used to generate QGS4 questions. This module generates questions that have an object property relation which is other than 'is-a', and 'is-an'. This approach can generate concept completion questions. An example of a question template for this category is as follows:

What is [P] of [X]?

The index is stored in relationArray[] as greater than 1. Therefore, the object property relation between [P] and [X] is determined using relMatrix[X][SC]. The algorithm for instantiating a question using the QGS4 is as follows:

```
if relMatrix[X][SC] is greater than 1
  then queArray.set(X, uniqueArray(idx_X)
       queArray.set(P, relArray(idx_P)
```

5.3 Analysis of Generated Questions

Four lists of questions were generated from AQGen for each data input, which includes a list of question from the individual class, two from terminology-based and one from a property-based approach. The total number of questions that can be generated from AQGen using ontoOS with 97 concepts and 16 question templates is 2103, while the total questions generated using ontoCN with 537 concepts and 17 question templates is 5326. The sample output for question generated from AQGen is presented in Figure 5.2 and Figure 5.3.

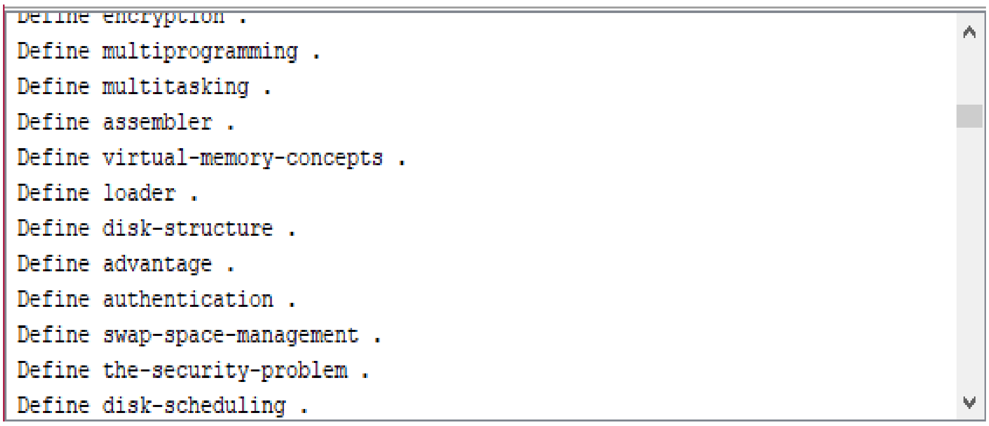


Figure 5.2: The sample questions generated by AQGen for QGS1 definition questions

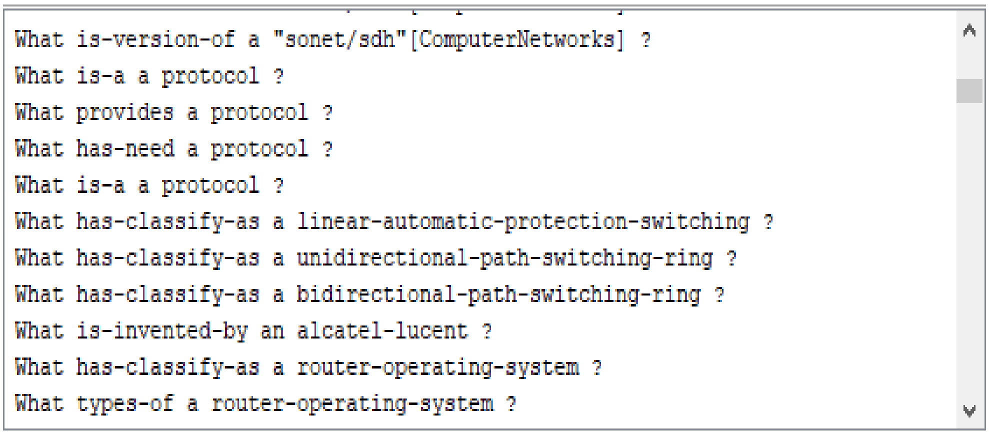

What is-version-of a "sonet/sdh"[ComputerNetworks] ?
What is-a a protocol ?
What provides a protocol ?
What has-need a protocol ?
What is-a a protocol ?
What has-classify-as a linear-automatic-protection-switching ?
What has-classify-as a unidirectional-path-switching-ring ?
What has-classify-as a bidirectional-path-switching-ring ?
What is-invented-by an alcatel-lucent ?
What has-classify-as a router-operating-system ?
What types-of a router-operating-system ?

Figure 5.3: The sample questions generated by AQGen for QGS4 concept completion questions

Table 5.2 shows the numbers of questions generated by each approach from two different ontologies which are Operating System (OS) and Computer Network & Security (CNS) using formula described in section 4.2.1 to 4.2.4.

Table 5.2: Total questions generated for each approach

Approach	Strategies	No. of Question Template	No of Question Generated	
			ontoOS	**ontoCN**
Individual class	QGS1	5	97 x 5 = 485	537 x 5 = 2685
Terminology-based	QGS2	5	86 x 5 = 430	433 x 5 = 2165
	QGS3	6	198 x 6 = 1188	73 x 6 = 438
Property based	QGS4	1	none	38 x 1 = 38

Table 5.3 shows the numbers of questions that are generated from each question template. OntoOS consist of 97 concepts, 86 hierarchical relations and no object property relations, while ontoCNS consists of 537 concepts, 433 hierarchical relation and 37 object property relations.

Table 5.3: Number of questions generated if only a question template for each question category.

Ontology	Number of questions generated by category			Total of Question Generated
	Definition	Concept Completion	Comparison	
ontoOS	Ind. Class: 97 Term: 86 Property: NA	Ind. Class: 97 Term: 86 Property: NA	Ind. Class: NA Term: 198 Property: NA	564
OntoCN	Ind. Class: 537 Term: 433 Property: NA	Ind. Class: 537 Term: 433 Property: 37	Ind. Class: NA Term: 73 Property: NA	2050

5.4 Summary

This chapter discussed the question generation tool AQGen that has been developed to generated questions for evaluation purposes. Three question generation approaches have been implemented inside AQGen to generate three question categories namely, definition, concept completion and comparison. 17 question templates were used as a stem that mapped to the concept in the ontology. The results obtained have been analysed, and the results are discussed in the next chapter.

Chapter 6

Methods and Metrics in Evaluation of Question Generation

6.1 Introduction

The AQGen has been developed to enable automatic generation of assessment questions using the proposed question generation strategies. The generated questions are analysed using a quantitative statistical method. The goals of this experiment are to answer research questions RQ3.

The research question RQ3 evaluates whether the combination of categorised question templates with QG strategies can generate acceptable assessment questions.

RQ3: Are the ontological question generation strategies able to generate acceptable assessment questions?

In order to answer this research question, we evaluate the suitability of the designed question generation strategies to map question templates to concepts in ontologies. This is obtained by evaluating the generated questions from AQGen. The acceptable generated questions were indicated by answering the following sub research questions:

RQ3.1: What is the agreement between experts for each QG strategy?

RQ3.2: To what extent do subject matter experts agree with the acceptability measures of AQGen generated questions?

RQ3.3: Are the generated questions grammatically correct?

RQ3.4: Are the generated question understandable?

RQ3.5: Are the generated questions relevant to the subject being evaluated?

RQ3.6: Can the generated questions be used as assessment questions?

RQ3.7: Do the generated questions using the correct question word?

The acceptability test was discussed in (Chali and Hasan, 2015) to measure relevancy and syntactic correctness of question generated using Name Entity (NE) information. The defined question is acceptable if the question shows no deficiency in terms of the criteria considered. For the purposes of this evaluation, another three measures are added as criteria for the question to be deemed as acceptable. An acceptable question for this evaluation is indicated using a set of Question Deficiencies (QDs) as presented in Table 6.1. In the case of unacceptable percentages recorded, the reasons that contribute to the unacceptable questions are analysed.

Table 6.1: Question Deficiencies

Question Deficiencies	Measure	Used in
QD1: This question is grammatically correct.	Syntactic Correctness	(Alsubait, et al., 2014)
QD2: This question is understandable	Ambiguity	(Olney, et al., 2012)
QD3: This question is within the context.	Relevance	(Olney, et al., 2012), (Alsubait, et al., 2014)
QD4: This question is useful.	Usefulness	(Alsubait, et al., 2014)
QD5: The question word is accurate.	Target type	(Olney, et al., 2012)

Next, we investigated which question strategy that would generate better assessment questions by carrying out an evaluation to find rating score differences between QG strategies in RQ3.8. The following research question was used as a guideline for the evaluation:

RQ3.8: Is there a difference in QDs rating between the question generation strategies?

Finally, we investigate whether the two ontologies used in this evaluation would affect the rating score for acceptability measures. RQ3.9 is set to provide an answer for this evaluation.

RQ3.9: Is there a difference in QDs rating between questions generated using ontoOS and ontoCN?

The next section in this chapter will discuss the experiment setup, the results obtained from the experiment and discussion.

6.2 Experiment setup

6.2.1 Instrument

A questionnaire was used as a medium for experts to evaluate the generated questions. The generated questions were randomly selected and manually copied into a questionnaire (see Appendix A) for the expert evaluation process. The questionnaire used for this evaluation has obtained approval from the university ethical committee, and the reference number of the approval is REGO-2017-2047. Four sets of questionnaires were created to reduce the effort for each expert – instead of evaluating hundreds of questions; each participant will only need to validate 50 to 70 questions. Table 6.2 shows the distribution of questions in each of the questionnaire sets.

Table 6.2: Distribution of questions in questionnaires

Set	Approach	Ontology	QG Strategies	No of Question	Total Questions
S1	Individual class	ontoCN	QGS1	Definition : 30 Concept Completion: 20	50
S2		ontoOS		Definition : 36 Concept Completion: 24	50
S3	Terminology-based	ontoCN	QGS2	Definition : 24 Concept Completion: 6	66
			QGS3	Comparison:36	
S4		ontoOS	QGS2	Definition : 24 Concept Completion: 6	66
			QGS3	Comparison:36	
S1	Property-based	ontoCN	QGS4	Concept Completion: 10	10
				TOTAL QUESTIONS	242

The questionnaire consisted of two sections. The first section required the experts to provide information on their backgrounds, such as institution name, department name, and their role either as student or lecturer. The second section presented a list of questions (see Appendix B) for evaluation. Each of the questionnaires differed in section 2, which is the list of questions to be evaluated. There are instructions on how to respond to each of the questions and detailed information on the question deficiencies, to help the participants to understand better. Participants were required to evaluate each question inside the questionnaire based on the five QDs and provide feedback in the form of the 3-point rating scale (Yes(1), No Option(2), No (3)). The rating 'YES' was to indicate if raters agreed with the indicated QD and 'NO' indicated not agree, whereas 'NO OPTION' indicated that raters were not sure about the answer. A further description of the 3-point scale is shown in Table 6.3.

Table 6.3: Question Deficiencies and its description

Question deficiencies(QD)	Descriptions
QD1: This question is grammatically correct.	If you believe the question is grammatically correct, please select **YES**. If you are not sure, please select **NO OPTION**. If you believe the question has a grammatical error(s), please select **NO** and state the reason.
QD2: This question is understandable.	If you understand the question, please select **YES**. If you are uncertain about the question, please select **NO OPTION**. If you do not understand due to the technical terms used, please select **NO** and state the reason.
QD3: This question is within the context.	If you believe the question is within the context of this subject, please select **YES**. If you are not sure, please select **NO OPTION**. If the question is not appropriate for this subject, please select **NO** and state the reason.
QD4: This question is useful.	If you believe the question is helpful in assessing student understanding, please select **YES**. If you are not sure, please select **NO OPTION**. If you believe the question is not helpful for assessment purposes, please select **NO** and state the reason.
QD5: The question word is accurate.	**The question word** is a function word used to ask a question, such as '*what*', '*explain*', and '*define*'. If you believe the question word is correct, please select **YES**. If you are not sure, please select **NO OPTION**. If you believe the question use wrong question word, please select **NO** and state the reason.

Questions generated from AQGen that needed to be evaluated might have grammatical errors, ambiguous meanings and sometimes might not even make sense at all. However, this would be the purpose of these questionnaires, namely, to ask participants to check for those deficiencies and later the results would indicate whether or not the strategies proposed to generate these questions were suitable.

6.2.2 Samples

Questionnaires were delivered to approximately 50 subject matter experts, including Computer Science lecturers/instructors and postgraduate students from several target universities, including Warwick, Northampton, and universities in Malaysia. Each questionnaire required approximately 5 participants. Table 6.4 shows the number of actual participants that have provided their feedback.

Table 6.4: Number of expected and actual participants

Set	Ontology	Expected number of raters	The actual number of raters
S1	ontoCN	5	5
S2	ontoOS	5	4
S3	ontoCN	5	3
S4	ontoOS	5	3

6.2.3 Measure

The questionnaire was used as a mechanism for evaluating the generated questions in this experiment. There were different measures calculated for this experiment, and these are discussed below:

A. Gwet C2 Inter-rater Agreement

Inter-rater reliability was calculated to measure agreement between experts on the generated questions on each question generation approach for both subjects. The agreement between raters was usually calculated using Fleiss' Kappa (Fleiss, 1971) reliability. However, the limitation of using kappa is that the low kappa value can be recorded even with a high value of the agreement (Wongpakaran et al., 2013; Gwet, 2008). Feedback from the questionnaire was observed, and it was found that the feedback given was of the high value of the agreement. Therefore, Gwet's Agreement Coefficient 2 (AC2) (Gwet, 2014) was used to calculate inter-rater agreement. Gwet's AC2 measure is calculated according to the following formula:

$$AC2 = \frac{p_a - p_\varepsilon}{1 - p_\varepsilon} \tag{6.1}$$

p_a = the proportion of observations in agreement
p_ε = the proportion in agreement due to chance

In this evaluation, we adopted different levels of agreement that are used in Landis and Koch (1977) as follow:

- <0 shows poor agreement
- 0.0 to 0.19 shows poor agreement
- 0.20 to 0.39 shows fair agreement
- 0.40 to 0.59 shows moderate agreement
- 0.60 to 0.79 shows substantial agreement
- 0.80 to 1.00 show almost perfect agreement

B. Percentage agreement

The percentage agreement was calculated to determine the percentage of experts that agreed with the feasibility measure for the generated question. 'Agree' in this context refers to annotated QDs with rating 'YES'. The result reports the percentage of the question at least one annotator annotated with ‘YES’ for all the QDs, and the percentage of questions when all annotators annotated with ‘YES’.

C. Means and Standard Deviation

Mean and standard deviation were calculated in this evaluation to measure the average rating score for the following groups:

i. each QG strategies
ii. each QDs

D. One-way ANOVA

One-way ANOVA was calculated to determine whether there are any statistically significant differences in question deficiencies between the four different QG strategies. We identify which QG strategies give better results for grammatical correctness, understandability, relevance, usefulness and appropriate question words for the generated questions. The results were reported using mean (M), standard deviation (SD), significant value (p) and F-test score. Therefore, the following research question, null hypothesis, and alternative hypothesis were posed:

Research question:

Is there a difference in QD rating between the question generation strategies?

Null hypothesis

There is no difference in QDs rating between QGS1, QGS2, QGS3, and QGS4.

Alternative hypothesis

There is a difference in QDs rating between QGS1, QGS2, QGS3, and QGS4.

The dependant and independent variables for this investigation are:

<u>Dependent variable</u>

QDs average score

<u>Independent variable</u>

QG Strategies: QGS1, QGS2, QGS3, and QGS4

Before we carried out one-way ANOVA, we ran Levene's test of equality of variance to test for homogeneity of variances. If the assumption of homogeneity of variances is met (significant value, $p > 0.05$), we report the F-test from ordinary one-way ANOVA; otherwise, we report F-test using Welch's ANOVA. The F-test is reported as follows:

$$F\ (df1,\ df2) = F\text{-}statistic,\ p = p\text{-}value$$

df1 indicates between groups degrees of freedom.

df2 indicates the within Groups degrees of freedom

F-statistic indicates the ratio of df1 and df2. When the null hypothesis is true, F-statistics is near to 1.

P-Value indicates the probability of obtaining the observed F-value if the null hypothesis is true.

Levene's test of equality of variance is also used to determine the type of post-hoc testing. Post-hoc testing was used as a follow-up analysis to determine which specific independent variable groups were significantly different from each other. Post-hoc testing was only carried out when there was a statistically significant difference reported by p-value ($p < 0.05$) in one-way ANOVA. There were two post hoc testing used in this evaluation which is Scheffe post hoc and Games-Howell post-hoc. If the assumption of homogeneity of variances was met, Scheffe post-hoc was carried out; otherwise, Games-Howell post-hoc was carried out.

E. Independence-samples t-test

The independent-samples t-test in this evaluation was used to determine if a difference existed between the means of two ontologies on a continuous mean rating for QDs. By applying the same QG strategies to both ontologies (OntoOS and ontoCN), we identified whether there was a significant difference in mean rating for QDs. We assumed the mean rating reported for both ontologies would have small differences because the concept name used in ontoOS was not specific as ontoCN. The result was reported using means (M), standard deviation (SD), significant value (p) and t-test. Therefore, the following research question, null hypothesis and alternative hypothesis were posed:

<u>Research question:</u>

Is there a difference in QDs rating between questions generated using ontoOS and ontoCN?

<u>Null hypothesis</u>

There is no difference in QDs rating between question generated using ontoOS and ontoCN.

<u>Alternative hypothesis</u>

There is a difference in QDs rating between question generated using ontoOS and ontoCN.

The dependant and independent variables for this investigation are:

<u>Dependent variable</u>

QDs average score

<u>Independent variable</u>

The two ontologies, which are ontoOS and ontoCN.

The Levene's test of equality of variance was carried out to test for homogeneity of variances. If the assumption of homogeneity of variances was met (significant value, $p > 0.05$), we reported the t-test from the 'standard' independent-samples t-test output in SPSS Statistics; otherwise we reported the t-test from a modified t-test that SPSS Statistics produced in its output, referred to as Welch t-test (Welch, 1947). The t-test is reported as follow:

$$t\ (df) = t\text{-}statistic,\ p = p\text{-}value$$

df indicates the degrees of freedom.

t-statistic indicates the difference relative to variation in sample data.

P-Value indicates the probability of obtaining the observed t-value if the null hypothesis is true.

6.2.4 Procedure

The generated questions from AQGen were copied into questionnaires for validation by experts. Experts, in this case, will refer to either lecturers/ students that teach/have learned Computer Science subjects.

The questionnaires were created on a Warwick web page and the link to the online questionnaire page, which is included inside the invitation letter, was sent to selected participants electronically beginning in August 2017. The questionnaire was accessible using the following link:

http://www2.warwick.ac.uk/fac/sci/dcs/research/edtech/surveys/aqg

Each participant was assigned a personal access code to access the questionnaire page, and this code was used in case of a withdrawal request later. The personal access code was provided in the invitation letter and is unique to each participant.

Invitation letters were sent through email to potential participants individually. Where contact details were unavailable, we requested our contacts to forward the invitation emails on our behalf. The potential participants were identified either through experts' teaching commitments found on university websites or through a recommendation of colleagues. Initially, we focused on participants from Computer Science since the useful ontologies we have access to are from the Computer Science domain. The participants did not have to understand how to use AQGen since the generated questions were copied into the questionnaires by the researcher manually. The potential participants were only required to validate the list of questions in the questionnaire. The initial subjects used to generate the questions in this work were core modules in Computer Science courses, and potential participants were likely to understand the questions being evaluated in the questionnaire. Since core modules in Computer Science were used, eliciting 20 responses for this evaluation was expected to be straightforward. Each questionnaire was designed to take approximately 30 minutes to complete.

If a participant chooses to participate in the study, by submitting a completed online questionnaire, the participant was deemed to have given their consent for the data that they supplied to be used anonymously in this study. The participants were given two weeks to return the complete survey, and if the completed survey was not received after that period, we assume that the participant was not interested. Submitting the online questionnaire marked the end of the participant's participation in the study.

6.3 Results of the Acceptability Measure for the Proposed QG Strategies

The results are discussed according to four question generation strategies and three question categories. The following subsection discusses the results obtained from statistical analysis using SPSS 24 software and Microsoft Excel. The discussion begins with inter-rater reliability measured using Gwet's AC2 in sub-section 6.3.1 and percentage of questions rated as acceptable by experts in section 6.3.2.

6.3.1 What is the agreement between experts for each of the QG Strategies?

This sub-section discusses the results obtained from the inter-rater agreement calculation for both subjects. Following sub-sections, A and B show the result of inter-rater agreement for both Computer Network and Operating System subject respectively. The inter-rater evaluation was conducted on four question generation strategies; QGS1, QGS2, QGS3, QGS4.

A. Computer Network (CNS) Subject

Table 6.5: QGS1 inter-rater agreement for CNS

Question deficiencies(QD)	Rater	AC2
QD1: This question is grammatically correct.	ABCDE	.784
QD2: This question is understandable.	ABCDE	.647
QD3: This question is within the context.	ABCDE	.476
QD4: This question is useful.	ABCDE	.594
QD5: The question word is accurate.	ABCDE	.548

Inter-rater reliability for QGS1 was calculated for the CNS subject on each of the five-question deficiencies QD1 to QD5. Gwet's AC2 was run to determine if there was an agreement between raters on whether 50 generated questions were exhibiting acceptable result. AC2 for the five raters is presented in Table 6.5. The agreement scores are above .6 for both QD1 and QD2, which is considered as being a substantial level of agreement between raters. However, the agreements reported using AC2 are above the moderate level for QD3, QD4, and QD5.

Table 6.6 QGS2 inter-rater agreement for CNS

Question deficiencies(QD)	Rater	AC2
QD1: This question is grammatically correct.	FGH	.963
QD2: This question is understandable.	FGH	.756
QD3: This question is within the context.	FGH	.925
QD4: This question is useful.	FGH	.588
QD5: The question word is accurate.	FGH	.681

Inter-rater reliability for QGS2 was calculated for the CNS subject on each of the five question deficiencies QD1 to QD5. Gwet's AC2 was run to determine if there was an agreement between raters on whether 66 generated questions were exhibiting acceptable result. AC2 for the three raters is presented in Table 6.6. Most of the agreement score is above 0.6, which is considered as the substantial level of agreement between raters except for QD4, which indicate a moderate level. The results also show a nearly perfect agreement for QD1.

Table 6.7: QGS3 inter-rater agreement for CNS

Question deficiencies(QD)	Rater	AC2
QD1: This question is grammatically correct.	FGH	1.000
QD2: This question is understandable.	FGH	.531
QD3: This question is within the context.	FGH	.344
QD4: This question is useful.	FGH	.406
QD5: The question word is accurate.	FGH	.359

Inter-rater reliability for QGS3 was calculated for the CNS subject on each of the five question deficiencies QD1 to QD5. Gwet's AC2 was run to determine if there was an agreement between raters on whether 66 generated questions were exhibiting acceptable result. AC2 for the three raters is presented in Table 6.7. The results show a perfect agreement for QD1. However, the rest of the QDs AC2 scores fell under the moderate and fair level of agreement.

Table 6.8: QGS4 inter-rater agreement for CNS

Question deficiencies(QD)	Rater	AC2
QD1: This question is grammatically correct.	ABCDE	.550
QD2: This question is understandable.	ABCDE	.415
QD3: This question is within the context.	ABCDE	.618
QD4: This question is useful.	ABCDE	.404
QD5: The question word is accurate.	ABCDE	.381

Inter-rater reliability for QGS4 was calculated for the CNS subject on each of the five question deficiencies QD1 to QD5. Gwet's AC2 was run to determine if there was an agreement between raters on whether 10 generated questions were exhibiting acceptable result. AC2 for the five raters is presented in Table 6.8. The agreement reported using AC2 is above the moderate level for all QDs, except QD5, which is considered as fair agreement.

B. Operating Systems (OS) Subject

Table 6.9: QGS1 inter-rater agreement for OS

Question deficiencies(QD)	Rater	AC2
QD1: This question is grammatically correct.	IJKL	.561
QD2: This question is understandable.	IJKL	.826
QD3: This question is within the context.	IJKL	.839
QD4: This question is useful.	IJKL	.833
QD5: The question word is accurate.	IJKL	.929

Inter-rater reliability for QGS1 was calculated for OS subjects on each of the five question deficiencies QD1 to QD5. Gwet's AC2 was run to determine if there was an agreement between raters on whether 50 generated questions were exhibiting acceptable result. AC2 for the four raters is presented in Table 6.9. Most of the agreement scores are above 0.8, which is considered as an almost perfect agreement between raters except for QD1, which is moderate agreement.

Table 6.10: QGS2 inter-rater agreement for OS

Question deficiencies(QD)	Rater	AC2
QD1: This question is grammatically correct.	MNO	.900
QD2: This question is understandable.	MNO	.950
QD3: This question is within the context.	MNO	.725
QD4: This question is useful.	MNO	.625
QD5: The question word is accurate.	MNO	.550

Inter-rater reliability for QGS2 was calculated OS subject on each of the five question deficiencies QD1 to QD5. Gwet's AC2 was run to determine if there was an agreement between raters on whether 66 generated questions were exhibiting acceptable result. AC2 for the three raters is presented in Table 6.10. The agreement reported using AC2 shows nearly perfect agreement for QD1 and QD2 but the moderate level for the rest.

Table 6.11: QGS3 inter-rater agreement for OS

Question deficiencies(QD)	Rater	AC2
QD1: This question is grammatically correct.	MNO	.979
QD2: This question is understandable.	MNO	.979
QD3: This question is within the context.	MNO	.917
QD4: This question is useful.	MNO	.896
QD5: The question word is accurate.	MNO	.896

Inter-rater reliability for QGS3 was calculated for OS subjects on each of the five question deficiencies QD1 to QD5. Gwet's AC2 was run to determine if there was an agreement between raters on whether 66 generated questions were exhibiting acceptable result. AC2 for the three raters is presented in Table 6.11. The agreement scores obtained show nearly perfect agreement across all QDs.

6.3.2 What are the percentages of subject matter experts that agree with the acceptability measure of AQGen generated questions?

The results in this section are based on the percentage of experts that agree with the acceptability measures for the generated question. The '%' symbol in Table 6.12 to Table 6.15 indicates the percentage of questions for at least one annotator annotated with 'YES' for all the QDs, whereas '%ALL' indicates the percentage of questions when all annotators annotated with 'YES'.

Table 6.12: Percentage of annotator annotated with rating 'YES' for QGS1

Question deficiencies(QD)	OS		CN	
	%	%ALL	%	%ALL
QD1: This question is grammatically correct.	100	60	98	34
QD2: This question is understandable.	100	40	100	72
QD3: This question is within the context.	100	26	100	74
QD4: This question is useful.	100	32	98	72
QD5: The question word is accurate.	100	24	100	90

Table 6.13: Percentage of annotator annotated with rating 'YES' for QGS2

Question deficiencies(QD)	OS		CN	
	%	%ALL	%	%ALL
QD1: This question is grammatically correct.	100	86.7	100	96.7
QD2: This question is understandable.	98	93.3	100	80.0
QD3: This question is within the context.	100	63.3	100	93.3
QD4: This question is useful.	90	50.0	100	63.3
QD5: The question word is accurate.	96	40.0	100	73.3

Table 6.14: Percentage of annotator annotated with rating 'YES' for QGS3

Question deficiencies(QD)	OS		CN	
	%	%ALL	%	%ALL
QD1: This question is grammatically correct.	100	97.2	100	100.0
QD2: This question is understandable.	98	88.9	100	58.3
QD3: This question is within the context.	100	86.9	100	41.7
QD4: This question is useful.	90	86.1	100	47.2
QD5: The question word is accurate.	96	86.1	100	44.4

Table 6.15: Percentage of annotator annotated with rating 'YES' for QGS4

Question deficiencies (QD)	CN	
	%	%ALL
QD1: This question is grammatically correct.	100	30
QD2: This question is understandable.	100	0
QD3: This question is within the context.	100	40
QD4: This question is useful.	100	10
QD5: The question word is accurate.	100	10

The percentage measures indicate that nearly all QDs in all sets of questionnaires for CN subject has at least one annotator with rating 'YES', except for QGS1 which shows 98 per cent, which indicates the missing data of 2 per cent. This result indicates that all questions have a certain level of acceptance by at least one annotator. On the other hand, OS subject shows that 10 per cent of the questions that no annotator gave rating 'YES' to the questions. The results indicate around 6 questions that have not been rated as acceptable. However, different percentages were reported when considering if all raters mark 'YES' in QDs. Most of the percentages or QGS1 in the Computer Network subject are higher than the Operating System subject except for grammar deficiency. The same pattern of percentages is reported for the QGS2 and QGS3. QGS4 is reported to have the lowest acceptability, with below 50 per cent for all QDs.

6.3.3 Are the generated questions grammatically correct?

Mean rating of the four question generation strategies for CNS subject are presented in Table 6.16. QD1 mean score increased from the QGS3($M = 1.02$, $SD = 0.11$) to the QGS2($M = 1.19$, $SD = 0.31$), QGS1($M = 1.19$, $SD = 0.34$, and QGS4($M = 1.54$, $SD = 0.51$).

Table 6.16: Mean rating for QD1 of Computer Network & Security subject

QG Strategy	Question Category	Mean	N	Std. Deviation
Strategy 1	Definition	1.2200	30	.41473
	Concept Completion	1.1600	20	.19029
	Total	1.1960	50	.34164
Strategy 2	Definition	1.2222	24	.33574
	Concept Completion	1.0556	6	.13608
	Total	1.1889	30	.31175
Strategy 3	Comparison	1.0185	36	.11111
	Total	1.0185	36	.11111
Strategy 4	Concept Completion	1.5400	10	.50816
	Total	1.5400	10	.50816
Total	Definition	1.2210	54	.37819
	Concept Completion	1.2481	36	.35191
	Comparison	1.0185	36	.11111
	Total	1.1709	126	.32904

Mean rating of the three question generation strategies for OS subject are presented in Table 6.17. QD1 mean score increased from the QGS3($M = 1.01$, $SD = 0.56$) to the QGS2($M = 1.08$, $SD = 0.26$), and QGS1($M = 1.51$, $SD = 0.47$, in that order.

Table 6.17: Mean rating for QD1 of Operating System subject

QG Strategy	Question Category	Mean	N	Std. Deviation
Strategy 1	Definition	1.6583	30	.47563
	Concept Completion	1.3083	20	.38341
	Total	1.5183	50	.46999
Strategy 2	Definition	1.0972	24	.28622
	Concept Completion	1.0000	6	.00000
	Total	1.0778	30	.25795
Strategy 3	Comparison	1.0093	36	.05556
	Total	1.0093	36	.05556
Total	Definition	1.4090	54	.48841
	Concept Completion	1.2372	26	.35955
	Comparison	1.0093	36	.05556
	Total	1.2464	116	.41111

6.3.4 Are the generated questions understandable?

Mean rating of the four question generation strategies for CNS subject are presented in Table 6.18. QD2 score increased from QGS3(M = 1.18, SD = 0.23) to the QGS2(M = 1.22, SD = 0.28), QGS1(M = 1.30, SD = 0.31) and QGS4(M = 1.60, SD = 0.38).

Table 6.18: Mean rating for QD2 of CNS subject

QG Strategy	Question Category	Mean	N	Std. Deviation
Strategy 1	Definition	1.2733	30	.35030
	Concept Completion	1.3500	20	.22361
	Total	1.3040	50	.30570
Strategy 2	Definition	1.2361	24	.28622
	Concept Completion	1.1667	6	.27889
	Total	1.2222	30	.28139
Strategy 3	Comparison	1.1759	36	.23212
	Total	1.1759	36	.23212
Strategy 4	Concept Completion	1.6000	10	.37712
	Total	1.6000	10	.37712
Total	Definition	1.2568	54	.32101
	Concept Completion	1.3889	36	.31127
	Comparison	1.1759	36	.23212
	Total	1.2714	126	.30430

Mean rating of the three question generation strategies for OS are presented in Table 6.19. QD2 mean score increased from the QGS3(M = 1.01, SD = 0.056) to the QGS2(M = 1.01, SD = 0.061), and QGS1(M = 1.13, SD = 0.24, in that order.

Table 6.19: Mean rating for QD2 of OS subject

QG Strategy	Question Category	Mean	N	Std. Deviation
Strategy 1	Definition	1.1750	30	.27189
	Concept Completion	1.0625	20	.15967
	Total	1.1300	50	.23819
Strategy 2	Definition	1.0139	24	.06804
	Concept Completion	1.0000	6	.00000
	Total	1.0111	30	.06086

Strategy 3	Comparison	1.0093	36	.05556
	Total	1.0093	36	.05556
Total	Definition	1.1034	54	.22133
	Concept Completion	1.0481	26	.14176
	Comparison	1.0093	36	.05556
	Total	1.0618	116	.17206

6.3.5 Are the generated questions relevant to the subject being evaluated?

Mean rating of the four question generation strategies CNS subject is presented in Table 6.20. QD3 mean score increased from the QGS2(M = 1.10, SD = 0.22) to the QGS3(M = 1.28, SD = 0.28), QGS4(M = 1.32, SD = 0.38, and QGS1(M = 1.58, SD = 0.48).

Table 6.20: Mean rating for QD3 of CNS subject

QG Strategy	Question Category	Mean	N	Std. Deviation
Strategy 1	Definition	1.6400	30	.55684
	Concept Completion	1.4975	20	.36218
	Total	1.5830	50	.48923
Strategy 2	Definition	1.0972	24	.20803
	Concept Completion	1.1111	6	.27217
	Total	1.1000	30	.21709
Strategy 3	Comparison	1.2778	36	.28172
	Total	1.2778	36	.28172
Strategy 4	Concept Completion	1.3200	10	.37947
	Total	1.3200	10	.37947
Total	Definition	1.3988	54	.51240
	Concept Completion	1.3838	36	.37432
	Comparison	1.2778	36	.28172
	Total	1.3599	126	.41897

Mean rating of the three questions generation strategies for OS subject are presented in Table 6.21. QD3 mean score increased from the strategy 3(M = 1.04, SD = 0.11) to the QGS1(M = 1.09, SD = 0.19), and QGS2(M = 1.11, SD = 0.16). in that order.

Table 6.21: Mean rating for QD3 of Operating System subject

QG Strategy	Question Category	Mean	N	Std. Deviation
Strategy 1	Definition	1.1250	30	.22505
	Concept Completion	1.0500	20	.10260
	Total	1.0950	50	.18824
Strategy 2	Definition	1.1389	24	.16787
	Concept Completion	1.0000	6	.00000
	Total	1.1111	30	.15982
Strategy 3	Comparison	1.0370	36	.10624
	Total	1.0370	36	.10624
Total	Definition	1.1312	54	.19998
	Concept Completion	1.0385	26	.09199
	Comparison	1.0370	36	.10624
	Total	1.0812	116	.16094

6.3.6 Are the generated questions can be used as assessment questions?

Mean rating of the four question generation strategies for CNS subject are presented in Table 6.22. QD4 mean score decreased from the QGS3(M = 1.26, SD = 0.24) to the QGS1(M = 1.34, SD = 0.38), QGS2(M = 1.39, SD = 0.36, and QGS4(M = 1.46, SD = 0.28), in that order.

Table 6.22: Mean rating for QD4 of CNS subject

QG Strategy	Question Category	Mean	N	Std. Deviation
Strategy 1	Definition	1.4133	30	.45466
	Concept Completion	1.2400	20	.20105
	Total	1.3440	50	.38128
Strategy 2	Definition	1.4167	24	.37105
	Concept Completion	1.2778	6	.32773
	Total	1.3889	30	.36181
Strategy 3	Comparison	1.2593	36	.24052
	Total	1.2593	36	.24052
Strategy 4	Concept Completion	1.4600	10	.28363
	Total	1.4600	10	.28363
Total	Definition	1.4148	54	.41576
	Concept Completion	1.3074	36	.25956

	Comparison	1.2593	36	.24052
	Total	1.3397	126	.33609

Mean rating of the four question generation strategies for OS subject are presented in Table 6.23. QD3 mean score increased from the QGS3(M = 1.05, SD = 0.12) to the QGS1(M = 1.10, SD = 0.19), and QGS2(M = 1.16, SD = 0.17).

Table 6.23: Mean rating for QD4 of OS subject

QG Strategy	Question Category	Mean	N	Std. Deviation
Strategy 1	Definition	1.1083	30	.20430
	Concept Completion	1.0875	20	.16771
	Total	1.1000	50	.18898
Strategy 2	Definition	1.1944	24	.16787
	Concept Completion	1.0000	6	.00000
	Total	1.1556	30	.16914
Strategy 3	Comparison	1.0463	36	.11691
	Total	1.0463	36	.11691
Total	Definition	1.1466	54	.19218
	Concept Completion	1.0673	26	.15096
	Comparison	1.0463	36	.11691
	Total	1.0977	116	.16821

6.3.7 Are the generated questions using the correct question word?

Mean rating of the four question generation strategies for CNS subject are presented in Table 6.24. QD5 mean score increased from the QGS3(M = 1.19, SD = 0.17) to the QGS2(M = 1.24, SD = 0.29), QGS1(M = 1.35, SD = 0.28, and QGS4(M = 1.74, SD = 0.23).

Table 6.24: Mean rating for QD5 of Computer Network & Security subject

QG Strategy	Question Category	Mean	N	Std. Deviation
Strategy 1	Definition	1.3400	30	.27867
	Concept Completion	1.3600	20	.30157
	Total	1.3480	50	.28517
Strategy 2	Definition	1.2847	24	.30880
	Concept Completion	1.0556	6	.13608
	Total	1.2389	30	.29583
Strategy 3	Comparison	1.1898	36	.17436
	Total	1.1898	36	.17436
Strategy 4	Concept Completion	1.7400	10	.23190
	Total	1.7400	10	.23190
Total	Definition	1.3154	54	.29093
	Concept Completion	1.4148	36	.34627
	Comparison	1.1898	36	.17436
	Total	1.3079	126	.29207

Mean rating of the three question generation strategies for OS subject are presented in Table 6.25. QD5 mean score increased from QGS3(M = 1.05, SD = 0.12) to QGS1(M = 1.08, SD = 0.25), and QGS2(M = 1.21, SD = 0.21).

Table 6.25: Mean rating for QD4 of Operating System subject

QG Strategy	Question Category	Mean	N	Std. Deviation
Strategy 1	Definition	1.1250	30	.31315
	Concept Completion	1.0000	20	.00000
	Total	1.0750	50	.24872
Strategy 2	Definition	1.2639	24	.19608
	Concept Completion	1.0000	6	.00000
	Total	1.2111	30	.20498
Strategy 3	Comparison	1.0463	36	.11691
	Total	1.0463	36	.11691
Total	Definition	1.1867	54	.27421
	Concept Completion	1.0000	26	.00000
	Comparison	1.0463	36	.11691
	Total	1.1013	116	.21333

6.4 Results of the Acceptability Measure Rating between QG Strategies

In the following subsection, we present a statistical test of significance for each subject between five question deficiencies. The test performed to identify which strategies that outperformed another in term of grammatical correctness, understandable, relevancy, usefulness and appropriate question word used. The results obtained answered RQ3.8.

6.4.1 Is there a difference in QD1 rating between the question generation strategies?

A one-way Welch ANOVA was conducted to determine if grammatically correct deficiency (QD1 mean score) was different across the four question generation strategies for questions generated from ontoCN. The effect size was calculated using Eta squared, $\eta^2 = .17$. Homogeneity of variances was violated, as assessed by Levene's Test of Homogeneity of Variance ($p = .001$). QD1 differed significantly across the four question generation strategies, Welch's $F(3,32.267)=8.638$, $p = .001$. Games-Howell post-hoc analysis revealed that the QGS3 had significantly less grammatical error compared to QGS1(0.17, 95% CI [0.04, 0.31), $p = .006$ and significantly less grammatical error than QGS4(0.52, 95% CI [0.02, 1.02), $p = .042$.

A one-way Welch ANOVA was conducted to determine if grammatically correct deficiency (QD1 mean score) was different across the three question generation strategies for question generated from ontoOS. The effect size calculated using Eta squared, $\eta^2 = .34$. Homogeneity of variances was violated, as assessed by Levene's Test of Homogeneity of Variance ($p = .0005$). QD1 differed significantly across the three question generation strategies, Welch's $F(2,51.518)=29.147$, $p = .0005$. Games-Howell post-hoc analysis revealed that the QGS3 had significantly less grammatical error compared to QGS1(0.51, 95% CI[0.35, 0.67]), $p = .0005$, and QGS2 had significantly less grammatical error than QGS1(0.44, 95% CI[0.25, 0.64]), $p = 0.005$).

6.4.2 Is there a difference in QD2 rating between the question generation strategies?

A one-way ANOVA was conducted to determine if understandable deficiency (QD2 mean score) for different QG strategies for questions generated from ontoCN. There was homogeneity of variances, as assessed by Levene's test of homogeneity of variances ($p = 0.26$). QD2 mean score was significantly differed across the four QG strategies, $F(3, 122) = 6.211$, $p = 0.01$, $\eta^2 = .13$. Scheffe post hoc analysis reveals that the QGS3 had significantly generated more understandable questions compared to QGS4(4.24, 95% CI[0.47, 0.80], $p = .027$, but no other group differences were statistically significant.

A one-way Welch ANOVA was conducted to determine if understandable deficiency (QD2 mean score) was different across the three question generation strategies for question generated from ontoOS. The effect size calculated using Eta squared, $\eta^2 = .12$. Homogeneity of variances was violated, as assessed by Levene's Test of Homogeneity of Variance ($p = .0005$). QD2 differed significantly across the three question generation strategies, Welch's $F(2,51.518)=29.147$, $p = .0005$. Games-Howell post-hoc analysis revealed that the QGS3 had significantly more understandable question compared to QGS1(0.12, 95% CI[0.04, 0.20]), $p = .003$, and QGS2 had significantly more understandable question compared to QGS1(0.12, 95% CI[0.03, 0.20]), $p = 0.04$).

6.4.3 Is there a difference in QD3 rating between the question generation strategies?

A one-way Welch ANOVA was conducted to determine if relevancy deficiency (QD3 mean score) was different across the four question generation strategies for question generated from ontoCN. The effect size calculated using Eta squared, $\eta^2 = .22$. Homogeneity of variances was violated, as assessed by Levene's Test of Homogeneity of Variance ($p = .0005$). QD3 differed significantly across the four question generation strategies, Welch's F(3,36.388)=12.259, p = .0005. Games-Howell post-hoc analysis revealed that the QGS2 had significantly better relevancy compared to QGS3(0.18, 95% CI [0.02, 0.34]), p = .026 and significantly better relevancy than QGS1(0.48, 95% CI [0.27, 0.69]), p = .0005. While QGS3 has significantly better relevancy compares to QGS1(0.31, 95% CI [0.09, 0.52]), p = .003.

A one-way Welch ANOVA was conducted to determine if relevancy deficiency (QD3 mean score) was different across the three question generation strategies for question generated from ontoOS. The effect size calculated using Eta squared, η^2 = .036. Homogeneity of variances was violated, as assessed by Levene's Test of Homogeneity of Variance (p = .023). QD3 mean score increased from the QGS3(M = 1.04, SD = 0.11) to the QGS1(M = 1.09, SD = 0.19), and QGS2(M = 1.11, SD = 0.16) in that order, but the differences between these QG strategies was not statistically significant, Welch's F(2, 67.917) = 3.096, p = .52.

6.4.4 Is there a difference in QD4 rating between the question generation strategies?

A one-way Welch ANOVA was conducted to determine if usefulness deficiency (QD4 mean score) was different across the four question generation strategies for question generated from ontoCN. The effect size calculated using Eta squared, η^2 = .032. Homogeneity of variances was violated, as assessed by Levene's Test of Homogeneity of Variance (p = .023). QD4 mean score decreased from the QGS3(M = 1.26, SD = 0.24) to the strategy 1(M = 1.34, SD = 0.38), QGS2(M = 1.39, SD = 0.36, and QGS4(M = 1.46, SD = 0.28), in that order, but the differences between these QG strategies was not statistically significant, Welch's F(3, 37.356) = 1.934, p = .14.

A one-way Welch ANOVA was conducted to determine if usefulness deficiency (QD4 mean score) was different across the three question generation strategies for question generated from ontoOS. The effect size calculated using Eta squared, $\eta^2 = .060$. Homogeneity of variances was violated, as assessed by Levene's Test of Homogeneity of Variance ($p = .0005$). QD4 differed significantly across the three question generation strategies, Welch's $F(2,68.259)=4.690$, $p = .012$. Games-Howell post-hoc analysis revealed that the QGS3 had significantly produced more useful question compared to QGS2(0.11, 95% CI [0.02, 0.19]), $p = .012$.

6.4.5 Is there a difference in QD5 rating between the question generation strategies?

A one-way Welch ANOVA was conducted to determine if question word deficiency (QD5 mean score) was different across the four question generation strategies for question generated from ontoCN. The effect size calculated using Eta squared, $\eta^2 = .24$. Homogeneity of variances was violated, as assessed by Levene's Test of Homogeneity of Variance ($p = .003$). QD5 differed significantly across the four question generation strategies, Welch's $F(3,36.473)=16.845$, $p = .0005$. Games-Howell post-hoc analysis revealed that the QGS2 had significantly better question word used compared to QGS4(0.50, 95% CI [0.25, 0.76]), $p = .0005$. The same pattern for QGS3 had significantly better question word used compared to strategy 1(0.16, 95% CI [0.03, 0.29]), $p = .011$ and significantly better question word used compared to QGS4(0.39, 95% CI [0.15, 0.63]), $p = .001$. While QGS1 has significantly better question word used compared to QGS (0.39, 95% CI [0.15, 0.63]), $p = 0.01$.

A one-way Welch ANOVA was conducted to determine if question word deficiency (QD5 mean score) was different across the three question generation strategies for question generated from ontoOS. The effect size calculated using Eta squared, $\eta^2 = .097$. Homogeneity of variances was violated, as assessed by Levene's Test of Homogeneity of Variance ($p = .024$). QD5 differed significantly across the three question generation strategies, Welch's $F(2,65.475)=7.566$, $p = .001$. Games-Howell post-hoc analysis revealed that the QGS1 had significantly better question word used compared to QGS2(0.14, 95% CI [0.01, 0.26]), $p = .027$. The same pattern for QGS3 had significantly better question word used compared to QGS1(0.16, 95% CI [0.06, 0.27]), p = .001.

6.5 Results of the Effect of QG Strategies rating score between Ontologies

The analyses in this section seek to identify whether our proposed QG strategies could be applied to a different domain ontology. The result obtained the answer to RQ3.9. This is done to analyse the effect of the feasibility score of three QG strategies on both ontoOS and ontoCS. Our guiding hypothesis in this evaluation was that the two ontologies would affect rating scores. Although both ontologies having the same conceptual structures, the labelling used to name the concept in the ontologies does not follow any standard. Labelling of concept name will affect the decision to determine the feasibility of the generated question. A Welch t-test was run to determine if there were differences in feasibility measure of generated questions for QGS1, QGS2 and QGS3 between ontoOS and ontoCS. Since ontoOS did not generate any QGS4 question, there was no comparison done for strategy 4. The test reports the acceptability comparison using significance value (p), Mean Difference (MD), Standard Error Different (SE) and t value.

6.5.1 Results of Acceptability Comparison between Ontologies for QGS1

The results for the acceptability score between ontologies for QGS1 are presented in table 6.26.

Table 6.26: Feasibility comparison between ontologies for QGS1

	QD1 Mean(s.d)	**QD2** Mean(s.d)	**QD3** Mean(s.d)	**QD4** Mean(s.d)	**QD5** Mean(s.d)
ontoCN	1.19(0.34)	1.3(0.31)	1.58(0.49)	1.34(0.38)	1.35(0.29)
ontoOS	1.51(0.47)	1.13(0.24)	1.09(0.19)	1.10(0.19)	1.08(0.25)
MD	-0.32	0.17	0.49	0.24	0.28
p	.0005	.002	.0005	.0005	.0005

The acceptability score for QD2 to QD5 had significantly better when using ontoOS compare to ontoCN. The ontoCN had significantly better for QD1 compared to ontoOS. The detailed interpretation of the results is as follows:

- There was a statistically significant difference in QD1 score between ontoCN and ontoOS, with ontoOS scoring higher than ontoCN, MD = -0.32, SE = 0.08, t(89.481) = 3.923, p = .0005.
- There was a statistically significant difference in QD2 score between ontoCN and ontoOS, with ontoCN scoring higher than ontoOS, MD = 0.17, SE = 0.05, t(92.473) = 3.175, p = .002.
- There was a statistically significant difference in QD3 score between ontoCN and ontoOS, with ontoCN scoring higher than ontoOS, MD = 0.49, SE = 0.07, t(63.197) = 6.583, p = .0005.
- There was a statistically significant difference in QD4 score between ontoCN and ontoOS, with ontoCN scoring higher than ontoOS, MD = 0.24, SE = 0.06, t(71.706) = 4.054, p = .0005.
- There was a statistically significant difference in QD5 score between ontoCN and ontoOS, with ontoCN scoring higher than ontoOS, MD = 0.28, SE = 0.05, t(96.223) = 5.102, p = .0005.

6.5.2 Results of Acceptability Comparison between Ontologies for QGS2

The results for the acceptability score between ontologies for QGS2 are presented in table 6.27.

Table 6.27: Feasibility comparison between ontologies for QGS2

	QD1 Mean(s.d)	QD2 Mean(s.d)	QD3 Mean(s.d)	QD4 Mean(s.d)	QD5 Mean(s.d)
ontoCN	1.19(0.31)	1.22(0.28)	1.10(0.22)	1.39(0.38)	1.24(0.29)
ontoOS	1.08(0.26)	1.01(0.06)	1.11(0.16)	1.16(0.19)	1.21(0.20)
MD	0.11	0.21	-0.11	0.23	0.03
p	.138	.0005	.822	.003	.674

The acceptability score for QD2 and QD4 had significantly better when using ontoOS compare to ontoCN. The mean score for QD1, QD3a and QD5 were not statistically significant. The detailed interpretation of the results is as follows:

- Although question generated using ontoOS(M=1.08) had less grammatical error as compared to ontoCN(M= 1.19, there was no statistically significant difference in QD1 score between ontoCN and ontoOS, MD = 0.11, SE = 0.07, t(56.036) = 1.504, p = .138. Therefore, the null hypothesis of equal population means between question generated from ontoCN and ontoOS for QD1 is accepted.
- There was a statistically significant difference in QD2 score between ontoCN and ontoOS, with ontoCN scoring higher than ontoOS, MD = 0.21, SE = 0.05, t(31.707) = 4.016, p = .0005.
- There was no statistically significant difference in QD3 score between ontoCN and ontoOS, MD = -0.11, SE = 0.05, t(53.298) = -.226, p = .822. Therefore, the null hypothesis of equal population means between question generated from ontoCN and ontoOS for QD3 is accepted.
- There was a statistically significant difference in QD4 score between ontoCN and ontoOS, with ontoCN scoring higher than ontoOS, MD = 0.23, SE = 0.07, t(41.097) = 3.200, p = .003.
- There was no statistically significant difference in QD5 score between ontoCN and ontoOS, MD = 0.03, SE = 0.07, t(51.630) = .423, p = .674. Therefore, the null hypothesis of equal population means between question generated from ontoCN and ontoOS for QD5 is accepted.

6.5.3 Results of Acceptability Comparison between Ontologies for QGS3

The results for the feasibility measure between ontologies for QGS3 are presented in table 6.28.

Table 6.28: Feasibility comparison between ontologies for QGS3

	QD1 Mean(s.d)	QD2 Mean(s.d)	QD3 Mean(s.d)	QD4 Mean(s.d)	QD5 Mean(s.d)
ontoCN	1.02(0.11)	1.18(0.23)	1.28(0.28)	1.26(0.24)	1.19(0.17)
ontoOS	1.01(0.06)	1.01(0.06)	1.04(0.11)	1.05(0.12)	1.05(0.15)
MD	0.01	0.17	0.24	0.21	0.14
p	.656	.0005	.0005	.0005	.0005

The acceptability score for QD2 to QD5 had significantly better when using ontoOS compare to ontoCN. The mean score for QD1 was not statistically significant. The detailed interpretation of the result is as follows:

- There was no statistically significant difference in QD1 score between ontoCN and ontoOS, MD = 0.01, SE = 0.02, t(70) = .447, p = .657. Therefore, the null hypothesis of equal population means between question generated from ontoCN and ontoOS for QD1 is accepted. Levene's Test for Equality of Variances is met.
- There was a statistically significant difference in QD2 score between ontoCN and ontoOS, with ontoCN scoring higher than ontoOS, MD = 0.17, SE = 0.04, t(38.997) = 4.190, p = .0005.
- There was a statistically significant difference in QD3 score between ontoCN and ontoOS, with ontoCN scoring higher than ontoOS, MD = 0.24, SE = 0.05, t(44.758) = 4.797, p = .0005.
- There was a statistically significant difference in QD3 score between ontoCN and ontoOS, with ontoOS scoring higher than ontoCN, MD = 0.21, SE = 0.04, t(50.665) = 4.778, p = .0005.
- There was a statistically significant difference in QD3 score between ontoCN and ontoOS, with ontoCN scoring higher than ontoOS, MD = 0.14, SE = 0.03, t(61.180) = 4.102, p = .0005.

6.6 Discussion

The experiment was conducted to evaluate the acceptability measure of the AQGen generated questions. The acceptability of the generated questions was measured using five question deficiencies which are syntactic correctness(QD1), understandability(QD2), relevancy(QD3), usefulness(QD4) and target type(QD5). The results interpreted in this chapter are according to research question RQ3. The discussion begins with the investigation to identify whether the question generated are rated as acceptable questions. Next, the analysis identifies if there is a difference in rating scores between question generation strategies. Lastly, the evaluation carried out to determine if there is a difference in an acceptability measure between question generated from ontoOS and ontoCN.

6.6.1 Inter-rater reliability

The analysis task begins with evaluating the agreement or consensus of generated questions against the five question deficiencies, between experts for each of the subjects using inter-rater reliability. The inter-rater agreement between raters was calculated using Gwet's AC2. Most of the AC2 score is above moderate level 0.5 for all question generation strategies in OS subjects, but low agreement on certain QDs reported on QGS4 and QGS3 for CNS subject.

The results from the percentage measures indicate that all questions have a certain level of acceptance by at least one annotator. However, there was a smaller percentage of acceptability recorded if all rater agreement is considered. The unacceptability of questions recorded was due to several reasons ranging from grammatical error, question not being understandable, question not within context, not useful, and wrong question word used.

6.6.2 Acceptability measure

The question generation strategies have been evaluated using two ontologies as a source of question keywords. Three question categories, namely definition, concept completion, and comparison, were generated from AQGen, and the acceptability of each QG strategies to construct this question categories was evaluated. The acceptability of the generated question is evaluated using five question deficiencies (QD1 to QD5), with a 3-point scale with 1-YES, 2-No option and 3-NO. Thus, the closer the mean score to 1, the better acceptability measure for that question.

Results revealed that total mean rating for all QDs was favoured to rating YES which indicates agreement to the QDs as presented in table 6.29. The results of this evaluation confirm that the QG strategies able to generate feasible questions.

Table 6.29: Overall mean rating for each QDs

	QD1 Mean(SD)	QD2 Mean(SD)	QD3 Mean(SD)	QD4 Mean(SD)	QD5 Mean(SD)
ontoCN	1.17(0.33)	1.27(0.30)	1.36(0.42)	1.34(0.34)	1.31(0.29)
ontoOS	1.25(0.41)	1.06(0.17)	1.08(0.16)	1.09(0.17)	1.10(0.21)

Although the results showed the proposed question strategies able to generate acceptable questions, there are still some questions that rated as less acceptable. Amongst the reasons are:

A. Grammatical error

The grammatical errors recorded as a missing or incorrect article (a, an) used in question sentence. The incorrect article used might lead to a wrong interpretation of the questions. It is, therefore, essential to have correct articles for question sentence. Although not many questions were rated as having an incorrect article, it does affects the QD1 rating score. The examples of these questions are “What is interrupt handler?” and “What is direct memory access?” that should be written as “What is **an** interrupt handler?” and “What is direct memory access?”

The results for OS subject shows mean rating, M = 1.66 in Table 6.17 indicate that the generated questions are favour toward getting the grammatical error. The grammatical error indicated from the expert’s feedback was inappropriately used or articles (a, an) in the question. The grammatical error was identified in the following questions:

- What is a secondary storage structure?
- What is a device driver?
- What is interrupt handler?

- What is direct memory access?
- What is disk scheduling?

Based on the questions above, the affected question template is of definition type, which is "*What is [X] ?"* and *What is [X] in [SC]?*. It is recorded that the grammatical error was due to the missing article on the generated questions. Compared to ontoCNS, the representation of the object properties follows the nouns it connected with. For example, when the nouns start with vowel letters such as "*input device*", the object property is set to "*is-an*" instead of "*is-a*". Therefore, the grammatical error about articles is resolved. Alternatively, the question template should be modified to capture information about articles such as "*What [art] [X] ?"*, with *[art]* represent the article. One simple way, to resolve this problem is by using a simple algorithm to detect the first string of *[X]*. If the first-string start with the letter *a, e, i, o,* or *u*, then variable *[art]* is set to 'is-an' otherwise set *[art]* to 'is-a'. Although the suggestion is not the best solution, exploring a better solution for correcting articles in the question sentence might be the interest for future work.

As the question generated is from question templates that we derived from the established and credential textbook revision questions and there are in a predefined structure; thus, we did not put checking grammar as the highest interest in the investigation. We invite experts from the computer science domain, who have experienced in learning or teaching Operating system and computer network subject. However, the result shows there is a small percentage of experts that evaluate the questions as having a grammatical error in Table 6.17 for OS subject. This is due to the background of this one expert that learned and taught in the non-English platform. Later, we sent the set of question to proof-reader to validate the accuracy of grammar used in this set of questions, and the outcome shows the questions is at appropriate grammar.

Another grammatical error is the used of a small letter for some concept name where it supposes to be capital letter such as TCP/IP.

B. The use of an abbreviation

On average, there is one expert for each set of questions that suggest writing the concept name in full rather than using abbreviation. Example of a question affected by this are:

- What uses an FTP?
- What is an application layer in the TCP/IP suite?
- Explain the term OSI Model.

C. Inappropriate question context

It is assumed that all questions would be in context since the key terms were extracted from course ontology. However, the results showed that several questions rated as not within the context of the subject being evaluated, such as and “Discuss what a presentation is?” and “Explain the term infrared”. This might be due to judging the question alone without considering the whole context of the questions set. Consider the following question setting:

TUTORIAL 1		TUTORIAL 2
Computer Network & Security		
Answer all questions.		Answer all questions.
1. Discuss what a presentation is.		1. Discuss what a presentation is.
2. Explain the term infrared		2. Explain the term infrared.

Example in Tutorial 1 would provide context for all questions with the subject title “Computer Network and Security” whereas the setting in Tutorial 2 does not indicate any context for the listed questions.

D. Question not useful

The result in Table 6.22 indicates that the definition questions for QGS1 generated not useful questions. The template is affected “What is [X]?”. The response from the respondent indicates that when the grammatical error and less relevant questions will be marked as not useful questions. The negative result here only appeared for CNS subject.

E. The inappropriate label of object properties

An inappropriate word used for an object property affected the meaning of the question. For example, the object property '*belongs-to'* in the triple '*802.11.g belongs-to a WLAN*' should be label as '*is-the-category-of'* and a question generated would be "What is the category of WLAN?" compared to 'What belongs to WLAN?' It will change the meaning of the sentence and hence affect the rating for QD1, QD2, QD3, and QD4. Therefore, appropriate names of object properties are as important as concept names to provide semantically correct representations.

F. Inappropriate used of question word

The result presented in Table 6.24 shows that the question word used in this strategy is 'What', which identified as inappropriate by 2 out of 5 experts. One of the experts suggested the following questions should use question word 'List' rather than asking 'What'.

What <u>is a type of</u> **Network Operating System**?

What <u>is a layer of</u> **TCP/IP**?

When the question word 'List' is used, it needs to indicate the number of items for each question asked. For example, **three (3)** in the following questions indicate the number of items required for that question:

List **three (3)** types of Network Operating System?

List **three (3)** layers of TCP/IP?

Another expert suggests using 'identify' instead of 'what'. Therefore, inappropriate question word used in these questions affect the rating score for QD5.

G. Ambiguous concept names

Result in Table 6.18 shows that some questions are said to be not understandable due to general keyword used, for example, “What is hybrid”. The keyword ‘hybrid’ is not giving any meaning to the question. There are only two questions that were recorded as not useful in the questionnaires, which are “What transmits a data?” and “What is meant by the term ssh-2 in ssh?”. The ssh-2 and data are individual, which might not suitable for definition question word while the first question was recorded to be a wide assumption or too general.

6.6.3 Mean rating comparison between question generation strategies

The analysis was conducted to identify which question generation strategies that could generate question with the highest acceptability score. Results in section 6.4.1 to 6.4.5 revealed that QGS3 produced more acceptable questions compare to other strategies for both ontologies across all QDs, except QD3 ($M = 1.28$, $SD = 0.28$) for question generated from ontoCN. This concluded that QGS3 generate less grammatical error question, more understandable question, more useful question and better question word used but less relevancy compared to QGS2 for question generated from ontoCN. QGS3 only generates a comparison type of questions. Therefore, we conclude that QGS3 can generate more feasible questions compared to other strategies.

On the other hand, QGS4 using ontoCN produced the less acceptable questions with mean rating closer to rating 2, $M = 1.54$(QD1), $M = 1.60$(QD2), $M = 1.32$(QD3), $M = 1.46$(QD4), and $M = 1.74$(QD5).

6.6.4 Mean rating comparison between ontologies

The proposed QG strategies were tested on two different ontologies to identify the generalisation of the proposed QG strategies when applied to different ontologies. The results in subsection 6.5.1 to subsection 6.5.3 revealed that generating questions using QGS1 on two different ontologies affect the acceptability score of generated questions. The results show that the generated questions using ontoOS had significantly better understandability, relevance, usefulness and question words used compared to question generated using ontoCN. The mean differences were small for understandability, usefulness and questions word which range between 0.17 and 0.28. However, mean different for relevancy is quite high where ontoOS mean shows very close to relevancy M = 1.09 but ontoCN means to show the generated questions nearly in between relevance and not relevance. The relevancy in this context checks whether the generated questions are within the context of the subject being tested, which in this case is Operating System and Computer Network & Security. Both of the ontologies show the same problem of relevancy.

On the other hand, there were also appeared to have meant different for QGS2 for all question deficiencies, but only QD2 and QD4 are statistically significant. The results revealed that generated question using QGS2 on ontoOS is significantly more understandable and generate more useful questions compare to ontoCN with MD = 0.21 and MD 0.23, respectively.

For QGS3, only QD1 was not statistically significant, but the rest were statistically significant differences for means between the two ontologies. The results revealed that ontoOS have better understandability, relevance, usefulness and question words used compared to ontoCN.

As a conclusion, the results revealed that there is a small effect on rating score onto both ontologies with ontoOS generates more acceptable questions compared to ontoCN for three QG strategies namely QGS1, QGS2, and QGS3.

6.7 Summary

This chapter discussed the results for the generated questions obtained to answer RQ3. The summary of the results is as follow:

1) There is a high inter-rater agreement of experts for OS subject (as presented in Table 6.9 to 6.11), but CNS subject appeared to be on a moderate level.

This might be due to the number of concepts in ontoCN were high for selection; which are 537 concepts compared to 97 concepts for OS subject.

2) Overall, the results presented in subsection 6.4.1 to 6.4.5 indicate that the question generation strategies able to generate acceptable questions with mean rating closes to 1.00. Results also reveal that QGS3 generate more acceptable questions compares to other strategies for both OS and CNS subject. The question template used in QGS3 is comparison question categories. Therefore, the strategy to used the sibling relationship in ontology to generate comparison questions give a valuable outcome for question generation research. Furthermore, QGS2 using hierarchical relationship has also shown an acceptable outcome. The question templates are comprising concept completion and definition categories. However, QGS4 shown a slightly less acceptable question generated. This due to inappropriate representation of object properties in ontoCN. The small questions selected for evaluation also affect the high mean rating for QGS4.

3) On the other hand, the evaluation to compare acceptability measure between ontologies has been conducted to identify whether the question generation strategies could be used in different domain. The result is presented in Table 6.29. The result shows that there are significant differences in the mean score. However, the differences are due to the inappropriateness of ontology representation and not dues to inappropriate keyword combinations in questions. Therefore, we conclude that the ontology could be used in both ontoOS and ontoCN. The ontoOS show better acceptability measure compared to ontoCN in term of understandability, relevance, usefulness and question word used.

Therefore, the result has suggested that ontology representation need to be redefined, and we have suggested a guideline for appropriate representation of ontology elements in section 8.2.1. The result also indicates that the categorised question templated proposed suitable for generating definition, concept completion and comparison question using knowledge from ontology. Finally, AQGen able to generate thousands of acceptable questions for assessment purposes.

In the next section, we will discuss the question selection process for filtering the AQGen generated questions. The filtering process is vital to generate feasible question sets for instructors and learners used.

Chapter 7

Updating AQGen with Question Selection

7.1 Introduction

In this chapter, we discussed the updated version of AQGen with question selection. The AQGen create an over-generation problem with a lengthy list of questions generated. To improve the feasibility of AQGen, we added simple question selection to allow easy selection of questions based on topic and question categories. We exploit the ontology structure to create topic-based indexing. The feasibility of the question selection measures the accuracy of the system output. Thus, in this chapter, we explained the process to answer RQ4.

The updated AQGen conceptual model with the question selection module is presented in Figure 7.1.

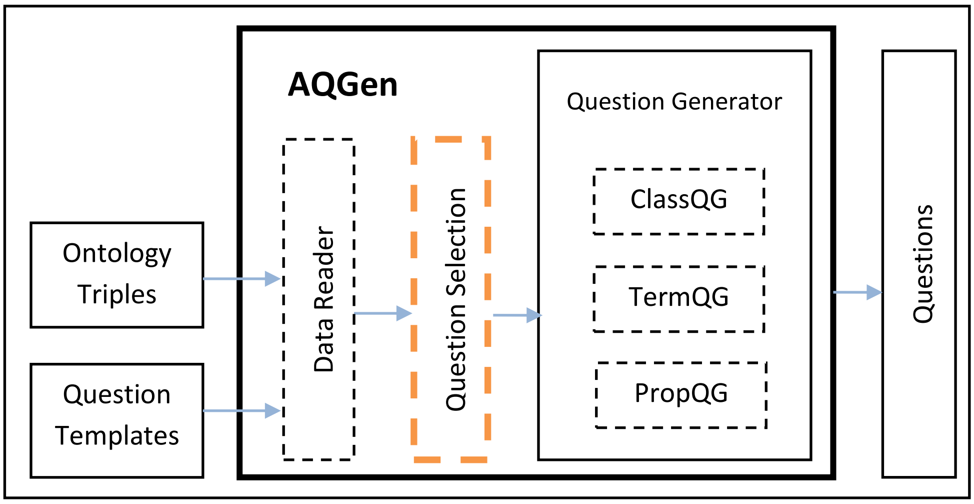


Figure 7.1: Updated AQGen conceptual model with question selection.

In section 7.2, we discuss the selection technique deployed in an existing e-learning system. In section 7.3, we continue with the explanation of the question selection module deployed in AQGen. In section 7.4, we discuss the results obtained.

7.2 Question Selection in e-Learning System

Assessment questions especially test and quizzes are generally given to students based on topics learned on the specific time frame. In the examination questions, the questions are usually designed and arranged according to topics. Thus, in this thesis, we designed the selection module based on topics. Topics are normally presented in the form of chapters of the syllabi.

Concerning question selection in e-learning, we took the selection approach that implemented in Modular Object-Oriented Dynamic Learning Environment (Moodle). Moodle is an open-source course management system has been developed by Martin Dougiamas and operation begins in 2002 (Dougiamas & Taylor, 2002; Dougiamas & Taylor, 2003). Moodle system allows the random selection of questions from question bank. Questions are manually inputted to Moodle by topics and user can select questions from each topic using the random function in Moodle.

Moreover, e-learning systems also using a randomized approach for question retrieval to allow dynamic question generation and reduced plagiarism amongst students (Lin & Chou, 2011).

Thus, in this thesis, we deployed a question generation module based on the topic selection using indexing and a randomized approach. We also implement topic-based indexing to represent topics.

7.3 Question Selection Module

In this stage, we enhance further the features of question generator to allow selection of the question, based on the topic. A long list of generated question in AQGen is filtered, and the selection of questions will be based on topic and question category. The selection approach will reduce the over-generation problem of questions where only required questions will be created. The selection modules deployed to demonstrate how the structured representation of the ontology could be used for filtering questions.

Topic indexing in this thesis refers to a technique of setting an index number to the ontology concepts. The idea of setting an index number is to increase the selection accuracy in any topic of choice. Index numbers are set to all concepts by traversing to all concepts according to the ontology structures. The indexing process begins with the top-level concept of ontology and moves from the left node to the right node. For the concept extraction process, the concept for the questions is selected using a random function. The process for the question selection module is shown in Figure 7.2.

Figure 7.2: Process flow for question selection module

The simplified Information Retrieval (IR) system discussed in (Soe, 2014) was adopted in our question selection module. The IR system contains four component, which is indexing, query formulating and analysing, retrieval and performance evaluation. We applied topic indexing, and random selection approaches for selecting questions. The following section discusses the four components in detail.

7.3.1 Indexing

Before the selection of question-based on topics taken place, a selection of the main topic must be made. Main topics for each subject were chosen from the main concepts in ontologies which are the top-level hierarchy of concepts, as shown in Figure 7.3 and Figure 7.4. There are nine top-level concepts in ontoOS ontology and eleven in ontoCN. These top-level concepts are used as the main topics for user selection.

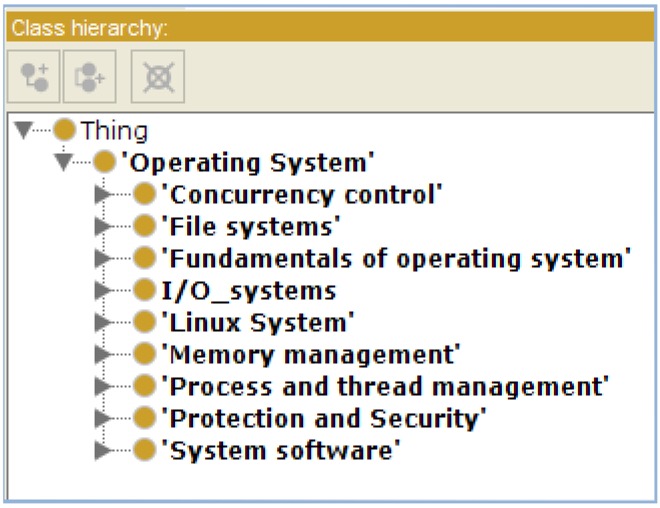


Figure 7.3: Snapshot from Protégé tool: Main topic selected for OS subject.

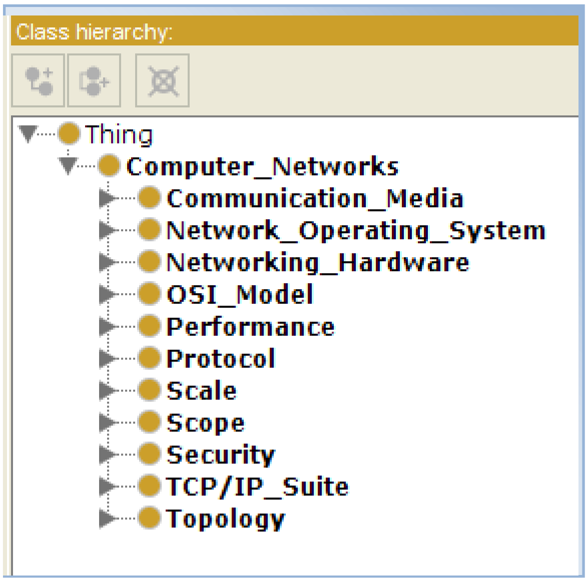


Figure 7.4: Snapshot from Protégé tool: Main topic selected for CNS subject.

Next, the index number was manually set to each of the main topics using numbers starting with number 0000 for the concept: ‘Operating System’ and

‘Computer Network’. The main concepts for each subject were numbered with 1000, 2000, and so on. Therefore, the topics had the index number as shown in Table 7.1 and Table 7.2

Table 7.1: Index number for OS main topics

OS Main Topics	Index number
Operating System	0000
Concurrency control	1000
File system	2000
Fundamental of operating system	3000
I/O systems	4000
Linux System	5000
Memory management	6000
Process and thread management	7000
Protection and security	8000
System software	9000

Table 7.2: Index number for CNS main topics

OS Main Topics	Index number
Computer Networks	0000
Communication media	1000
Network Operating System	2000
Networking Hardware	3000
OSI Model	4000
Performance	5000
Protocol	6000
Scale	7000
Scope	8000
Security	9000
TCP/IP Suite	10000
Topology	11000

The subtopics were also indexed with a continuous number from its main topic. For example ‘coaxial cable’ and ‘wired/Guided Media’ are the subconcepts of ‘Computer Network’, therefore the subconcepts were given an index number 1001 and 1002 consecutively as shown in Figure7.5. The first row in the figure indicates the taxonomy level of the concept in the ontology.

```
0 0000 Computer_Networks;
1 1000 Communication_Media;
2 1001 "Wired/Guided_Media";
3 1002 Coaxial_Cable;
4 1003 Biaxial;
4 1004 Hardline;
4 1005 Oliver_Heaviside;
4 1006 Radiating;
4 1007 Radio_Frequency_Signals;
```

Figure 7.5: Sample of the subtopic index number.

7.3.2 Query formulation and analysing

The question selection process begins with the selection of the subject. The main topics for the subject were listed in the dropdown list box as shown in Figure 7.6. Once the topic selected, the next step is to select the question type which indicates question taxonomies, as shown in Figure 7.7. Next step is to click 'ADD' button to add the selection to the question requirement list as shown in Figure 7.8.

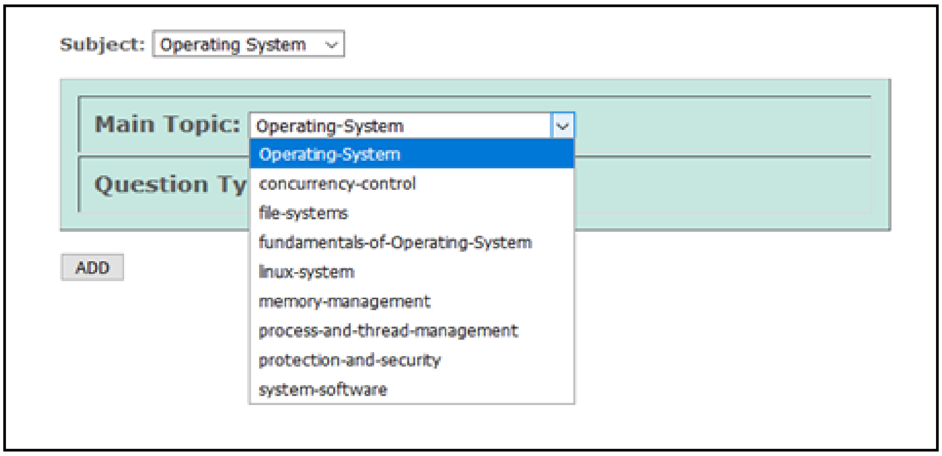


Figure 7.6: Main topic selection.

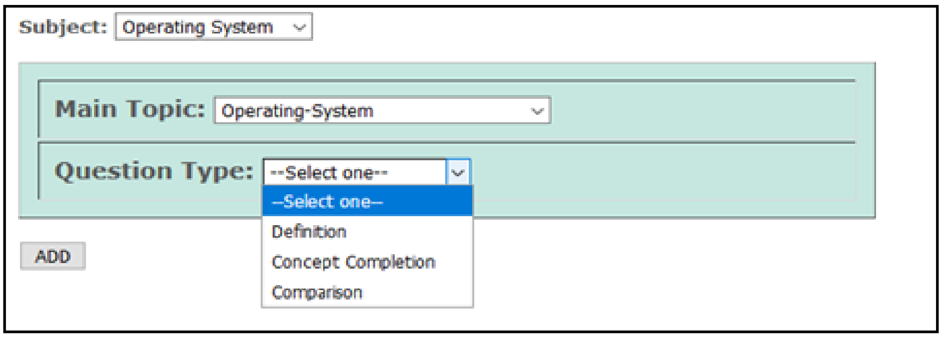


Figure 7.7: Question taxonomy selection.

Subject	Topic	Category	Remove
OS	Operating-System	Definition	X
OS	concurrency-control	Definition	X
OS	file-systems	Comparison	X
OS	concurrency-control	Comparison	X
OS	memory-management	Concept Completion	X
OS	protection-and-security	Concept Completion	X

Generate Questions

Figure 7.8: List of selection requirement.

After the selection process complete, by clicking 'Generate Question' button will generate questions as presented in Figure 7.9.

QUESTIONS
1. What does it means by "I/O systems" ?
2. What is meant by the term secondary-storage-structure in "I/O-systems" ?
3. Define deadlock .
4. What is detection in deadlock ?
5. Compare between directory-systems and file-support .
6. With regards to concurrency-control , explain the difference between a deadlock and software-and-hardware-solutions .
7. Explain the term .
8. Within the context of , explain the term [X] .
9. Discuss what is protection and security .
10. Within the context of protection-and-security , explain the term security .

Figure 7.9: List of generated questions.

7.3.3 Question retrieval

Question extraction is the backend process for generating question based on the selection in Section 7.3.2. The index number was used to retrieve concepts together with random selection run for choosing sub-topic. For example, if the main topic 'Concurrency control' was selected, the question generator read the index as 1000. In order to choose the subtopic, the question selector produces a random number between 0 and the number of subclass for

‘Concurrency control’. Let say, ‘Concurrency control’ has 10 subclasses; the random number is between 0 and 10. The random number produced, added up to the main topic index to find the match subtopic index. For example, if the random number is 3, the selected concepts would be of index 1003. Therefore, questions would be generated using the concepts of index 1003.

7.3.4 Evaluation

In this section, we discussed the evaluation method by explaining the instrument and the procedure used to conduct the experiment and measure used to indicate the accuracy of proposed topic-based indexing and randomised approach for question selection.

A. Instrument

The evaluation for the feasibility measure of question selection module is measured using both OS and CNS subjects. The evaluation determines the feasibility based on whether the intended questions are generated as a complete question without any missing concepts. Four topics from each subject are selected, and the number of question evaluated are presented in Table 7.3 and Table 7.4. This dataset is queried for the question selection module.

Table 7.3: Queries used for evaluating feasibility for OS.

Topics	Number of questions evaluated
Topic 1: Concurrency control	Definition: 40, Concept Completion: 40, Comparison: 20
Topic 2: File system	Definition: 40, Concept Completion: 40, Comparison: 20
Topic 3: Memory Management	Definition: 40, Concept Completion: 40, Comparison: 20
Topic 4: Linux system	Definition: 40, Concept Completion: 40, Comparison: 20

Table 7.4: Queries used for evaluating feasibility for CNS.

Topics	Number of questions evaluated
Topic 5: Communication media	Definition: 40, Concept Completion: 60, Comparison: 30
Topic 6: OSI Model	Definition: 40, Concept Completion: 60, Comparison: 30
Topic 7: Security	Definition: 40, Concept Completion: 60, Comparison: 30
Topic 8: Networking hardware	Definition: 40, Concept Completion: 60, Comparison: 30

B. Procedure:

The feasibility evaluation of question selection model is based on the output of updated AQGen. The questions were selected according to the queries presented in Table 7.2 and 7.3, and there were four topics for each subject. The questions were then analysed to identify all questions variables were instantiated. The number of correctly instantiated questions was recorded, and calculation for precision, recall and accuracy was performed using MS Excel tool. The formula and classification of the questions for evaluation are discussed in the next section.

C. Measure:

The evaluation used to identify the accuracy of topic-based question selection process using accuracy, precision and recall. Accuracy measures the performance of the proposed topic-based indexing to filter questions in order to obtained high feasibility of AQGen. Precision is used to identify how precise the selection algorithm, thus the higher the precision, the less irrelevant question retrieved. While recall measures how complete the selection is, thus the higher the recall, the less missing questions. In order to calculate the precision and recall, the system outputs are further breakdown into the following categories:

True Positive (TP): System is classified as will generate questions and had, in fact, generated the questions.

False Positive (FP): System is classified as will generate questions but did not generate the questions.

True Negative (TN): System is classified as will not generate questions and in fact did not generate the questions.

False Negative (FN): System is classified as will not generate questions but actually generated the questions.

Thus, precision and recall are calculated as follows:

$$\text{Precision, P} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{6.1}$$

$$\text{Recall, R} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{6.2}$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \tag{6.3}$$

$$\text{Average precision} = \frac{\sum_1^n \text{Precision}}{\text{Topic count}} \tag{6.4}$$

7.4 Result and Discussion

The results for precision, recall, and accuracy is presented in Table 7.5 and Table 7.10. The results indicated in this section used to determine how feasible the topic indexing in filtering questions. The precision, recall, and accuracy was calculated using formula 6.1 to 6.4 in section 7.3.4. The results presented indicate that our topic indexing with random selection is approximately 98 per cent accurate for OS queries and 99 per cent accurate for CNS queries. The results also report a perfect recall score of 1.00 for which means all intended questions were retrieved from the question selection for both subjects. On the other hand, the results also indicate that more than

80 percents questions retrieved were as expected, with the average precision obtained were 0.821 for OS and 0.919 for CNS. Thus, overall results indicate the updated AQGen is feasible of generating intended question using topic-based indexing with random selection.

A. Operating Systems

Table 7.5: OS - Precision, recall and accuracy for definition questions

	TP	FP	TN	FN	Precision	Recall	Accuracy
Topic 1	20	0	183	0	1.00	1.00	1.000
Topic 2	20	0	183	0	1.00	1.00	1.000
Topic 3	15	5	178	0	0.75	1.00	0.975
Topic 4	12	8	175	0	0.60	1.00	0.969

Table 7.6: OS - Precision, recall and accuracy for concept completion questions

	TP	FP	TN	FN	Precision	Recall	Accuracy
Topic 1	20	0	183	0	1.00	1.00	1.000
Topic 2	20	0	183	0	1.00	1.00	10.00
Topic 3	12	8	171	0	0.60	1.00	0.958
Topic 4	15	5	168	0	0.75	1.00	0.973

Table 7.7: OS - Precision, recall and accuracy for comparison questions

	TP	FP	TN	FN	Precision	Recall	Accuracy
Topic 1	20	0	198	0	1.00	1.00	1.000
Topic 2	20	0	198	0	1.00	1.00	1.000
Topic 3	13	7	185	0	0.65	1.00	0.966
Topic 4	10	10	188	0	0.50	1.00	0.952

B. Computer Network & Security

Table 7.8: CNS - Precision, recall and accuracy for definition questions

	TP	FP	TN	FN	Precision	Recall	Accuracy
Topic 5	39	1	931	0	0.98	1.00	0.999
Topic 6	40	0	930	0	1.00	1.00	1.000
Topic 7	40	0	930	0	1.00	1.00	1.000
Topic 8	40	0	930	0	1.00	1.00	1.000

Table 7.9: CNS - Precision, recall and accuracy for concept completion questions

	TP	FP	TN	FN	Precision	Recall	Accuracy
Topic 5	54	6	1478	0	0.90	1.00	0.996
Topic 6	54	6	1478	0	0.90	1.00	0.996
Topic 7	50	10	1482	0	0.83	1.00	0.994
Topic 8	45	15	1487	0	0.75	1.00	0.990

Table 7.10: CNS - Precision, recall and accuracy for comparison questions

	TP	FP	TN	FN	Precision	Recall	Accuracy
Topic 5	26	4	1162	0	0.87	1.00	0.997
Topic 6	30	0	1158	0	1.00	1.00	1.000
Topic 7	30	0	1158	0	1.00	1.00	1.000
Topic 8	24	6	1164	0	0.80	1.00	0.995

The high accuracy obtained indicates that the topic-based indexing structure that exploited ontology structure to match concepts can filter questions correctly. The proposed topic indexing techniques could avoid complex processing in text processing, and as compared to the text-similarity techniques, it is difficult to obtained relationship information between concepts. With the ontology structure, the sub-topic is easily retrieved from the ontology taxonomy. The experiment should obtain a result of 100 per cent accuracy, which indicates that the topic indexing able to generate questions based on the chosen topic accordingly. However, accuracy is slightly less than 100 per cent due to missing concepts appeared when the question generated. The missing concepts appear because of some topic that did not have a child concept. Therefore, the evaluation to identify whether the question falls under the correct topic cannot be established. Therefore, the accuracy in this result affected by the missing information on the question, but not the evaluation of the topic to sub-topic relation.

7.5 Summary

In this chapter, we proposed a question selection module to improved AQGen feasibility in term of filtering intended question from a lengthy list of questions. Simple IR system has been implemented by adopting topic-based indexing to represent concepts and random selection for question retrieval. The results obtained has indicated that the proposed technique of question selection using topic-based indexing can improve AQGen feasibility.

AQGen generated a tremendous amount of questions, thus required some question filtering to generate a question set that is useful to instructors and learners. Manual selection from thousands of questions generated is not a feasible task. It is essential to have question filtering, so only the required question set presented to instructors and learners. It is part of the contribution as most of the automatic question generator produces a tremendous amount of questions, and it needs to be filtered out when instructors need to create a question set. As the evaluated ontology (ontoOS and ontoCNS) does not have information/knowledge about topic chapters, the topics indexing is introduced as a way to represent knowledge about topic chapters in the ontology. The small-scale experiment conducted shown that topic indexing could filter the question to generate the required question set without any complicated process. The contribution of this chapter lies in the way we arrange and ordering the indexing for concepts in the ontology. We exploit the hierarchical relationship for classifying the main topic with sub-topic and randomised the selection of subtopic for question generation.

Chapter 8

Conclusions and Recommendations

8.1 Introduction

In the preceding chapters, we discussed the related work in automatic question generation (Chapter 2), we explored how well the ontologies can be used as an input source for generating assessment question (Chapter 3), we described how to designed categorized question templates and question generation discussed, we described an implemented AQGen tool to apply the proposed question generation strategies and categorised question templates (Chapter 5), we evaluate the generated assessment questions (Chapter 6), and we presented and updated AQGen with simple topic-based question selection technique using topic-based indexing and randomised function (Chapter 7). In this chapter, we conclude by summarizing our research findings, discussing the recommendation for future work, and reiterating our contributions.

8.2 Discussion of the Findings

This section discusses three interesting issues which contribute to existing theories: use of course ontology for automatic question generation; ontological approach for automatic question generation; topic-based question selection to increase the feasibility of the tool.

8.2.1 Use of course ontology for automatic question generation

The purpose of using course ontology in this research on automatic question generation was to benefit from the structure of ontology representation for education assessment questions. Although the structured representation of ontology limits the generation of complex questions, the ontology representation allows difference question categories to be created, as discussed in Section 2.1.3. The results about RQ1, presented in Section 3.4, show difference question categories; definition, concept completion, and comparison could be generated with appropriate question templates. The results also indicate that most of the concepts in the ontology are suitable as a source for keywords in assessment questions. However, a few concepts were labelled with inappropriate names was found, and it is indicated in Section 3.4.5 that would generate incorrect questions. This is consistent with Papasalourus et al., (2008) suggestion for input ontology to adhere to certain conventions in order to generate a syntactically correct sentence.

An analysis to investigate how well the ontology can be used for generating assessment questions has been presented in Chapter 3. The experiment aims to evaluate the suitability of the ontoOS for question generation. Suitability, in this case, refers to the appropriateness of concepts represented in the ontology. This is done through validation of ontoOS concepts with keywords exist in the OS textbook. Question competency approach has been adopted, where revision questions from the OS textbook are used as a competency question for the validation process. In this process, the similarity between revision questions' keyword is compared with ontology concepts. The similarity evaluation using the Sorensens Similarity index is discussed in section 3.3. The result has shown that half of the concepts in ontology matched keyword in assessment questions. This result has been made as an indicator to use ontoOS as a source of knowledge for AQGen.

On top of that, another interesting finding has been discovered from this experiment which is a guidance of ontology representation for question generation. From the experiment, it is found that only half of the concepts in the ontology that matched the keywords from assessment questions. Further investigation had been undertaken to analyse the unmatched concepts in the ontology. The analysis has identified 4 improper representation of concepts' name in ontoOS as presented in section 3.4.5, and the suggestion for proper naming is discussed here as a guideline for future ontology engineer.

First, long concept name used in naming concept such as "*interaction-between-processes-and-os-management*". The representation includes too many information and could possibly be challenging to run the reasoning process. Generating question from this concept will introduce ambiguity or need dedicated question template for this concept such as "Do there has any *interaction between processes and os management*". However, this question template pattern does not match other concepts in the ontoOS. The long name provided here could be split into 3 ontology elements i) concept: *processes*, ii) concept: *os-management*, iii) object property: *have-interaction*.

Second, a concept naming used more than two nouns such as "*os-services-and-components*". A concept name should use single nouns as suggested in (Bergman, 2010) as a best practice. In this example, it is not clear whether both "*os-services*" and "*component*" as one entity or separated entity. When using this concept for question generation, it creates ambiguity to the question. This work suggests the use of single nouns for naming concept, and if a noun contains two words such as Operating System, it could be written as "*operatingSystem*" with camel case for the first letter of the second word.

Third, concept name repetition in sub-class and super-class. For example, the representation of "*goal-of-protection is-a protection*" has created a redundant word in the following question:

"Explain the goals of protection in protection."

The representation might be suitable for other computer problem but it clearly not suitable for question generation, especially for the above question stem.

Finally, action verb as part of concept name such as "*implementation-of-access-matrix*". Typically, a concept use nouns or nouns phrase for naming and verb action is used to represent object property. The word "*implementation-of*" should be an object property, and "*access-matrix*" is a concept name.

Therefore, this contribution of this finding is a guideline for proper ontology concepts representation for future ontology engineer to consider when developing an ontology, especially for question generation. this finding would benefit those who develop course ontology.

The process of identifying question generation strategies, as described in Chapter 4, produced four question generation strategies. The identification process began by analysing how keywords in a question are presented in ontologies and the relation between them. This approach complies with Papasalouros et al. (2008) ontological approach to question generation for generating distractors of MCQ. The question generation strategies were thus designed to map categorised question templates with concepts in the ontology. The first question generation strategy used any concepts that exist in the ontology, while the rest used relations information to select appropriate concepts in ontologies. The strategies were also determined by the structure of the question templates. Therefore, a list of question generation strategies and its questions templates were proposed in Section 4.4.

8.2.2 Ontological approach for automatic question generation

Previous studies have had calculated inter-rater agreement using Cohen's inter-rater reliability (Cohen, 1960) for two raters and Fleiss inter-rater reliability (Fleiss, 1971) for more than two raters. In our evaluation reported in Section 6.2.1, the inter-rater agreement was calculated using Gwet's AC2. As far as we are concerned, there is no official results appear to have been released for calculating the inter-raters agreements using Gwet's AC2; thus, comparison with the previous literature is somewhat limited. Gwet's AC2 was used due to high agreement of raters found in the dataset. The results indicate that there was a high agreement for OS subjects with almost all rated as close to a near-perfect agreement, as compared to CNS subject, which indicates a more moderate level of agreement.

Overall, most mean ratings reported for QDs in Table 6.23 to 6.27 support the conclusion that our question generation strategies are able to generate acceptable questions, with the mean rating range between M = 1.00 and M = 1.36, the closer the mean value to 1, indicate less deficiency the question is.

The comparison between question generation strategies in Section 6.4 and 6.5 is the most interesting comparison in our evaluation. The first main finding of the question generation strategy comparison with regard to RQ3.8 is that QGS3 using terminology-based approach generates more acceptable questions and the less acceptable questions were generated by QGS4 that using property-based approach for CNS subject. This is consistent with Papasalourus et al., (2008) analysis, which found that questions generated using property-based are very difficult to manipulate compared to terminology-based. QGS3 was also reported as a strategy that generated the most acceptable question for OS subject; while QGS2 was reported as generating less relevant and understandable questions, compared to QGS1. The second finding was a comparison of generated questions between two ontologies; ontoOS and ontoCN. The results reported with regard to RQ3.9 indicate that there is a significant difference between questions generated ontoOS and ontoCN, where the results can be interpreted as question generated using ontoOS has better opportunity to generate acceptable questions. The results also proved why there was low agreement reported in section 6.2.1 for CNS subject. One of the reasons might be due to the number of concepts represented in the ontoCN (537 concepts) being big as compared to ontoOS (97 concept). The higher number of concepts might increase the chances of choosing inappropriate concept names.

8.2.3 Categorized question templates

Question templates are used in this work as a general question stem with placeholders that could be filled in to generate the same categories of a question with different keywords from course ontology. Leo et al. (2019) defined the question template as:

“A generic skeleton of a question with place-holders that can be filled in with relevant question content to make various questions of a similar type.”

Revision question of OS subject with 15 chapters and 256 questions were analyzed to identify question categories used in this subject. The analysis identified 5 major categories exist in the revision questions which the highest is concept completion, followed by verification and features specification. There other two were definition and comparison question. Out of 5, concept completion, definition and comparison were chosen to be implemented in this work due to the structure of the question that makes it possible to be generated using an ontology, and to the best of our knowledge, it has not been implemented in previous research. Based on these 3 question categories, 17 questions proposed in this work as shown in Table 4.1. Out of the 17 question templates, 7 are of definition type, 4 are of concept completion type, and 6 are of comparison type.

The 17 question templates which classified in 3 question categories is one of the major contributions of this work. Many previous works have suggested different question templates, especially for MCQ, and each will fit the purpose of their domain ontology. To the best of our knowledge, there are no question templates that suitable for all domain. We have to prove this in the experiment which presented in (Ibrahim Teo & Joy, 2016) where we implement our suggested question templates into Travel ontology. The result shows that the question templates are not suitable for the travel domain as the nature of question asking in travel domain did not ask about definition question such as “Define Sydney” is deemed as an inappropriate question. This is also supported by the research in (Leo et al., 2019) which design the question templates that only suitable for questions in the medical domain as presented in section 2.4.2.

Therefore, this work has added up the collection of existing question templates for ontology-based question generation to prepare for emerging technologies in e-learning, especially online assessment. The proposed question templates can be used with ontology for a domain that needs assessment questions in these 3 question categories (concept completed, comparison and definition).

8.2.4 Topic-based indexing for question selection

The question selection module was developed to improve the feasibility of AQGen in term of question filtering. The result obtained from section 7.4 indicates that the topic-based indexing able to filter questions and returned intended results.

8.3 Summary of Contributions

This research will significantly contribute to enhancing expertise in education sectors, particularly in the area of education assessment quality and process improvement. This research will also broaden the perspective of automatic question generation techniques and algorithms in the area of education, and particularly in the assessment. It could provide support and intelligent assistance to the instructor to generate question automatically. It could also indirectly increase staff time allocated to developing new skills and construct valuable feedback for students rather than spending their time on question creation.

The primary contributions of this research are divided into three key areas:

8.3.1 Contribution to the understanding of question generation using an ontology element

This thesis contributes to an understanding of data input preparation for generating questions automatically using ontology. The pre-analysis described in Chapter 3 was designed to answer RQ1, which contributes by using assessment questions to validate ontology, in order to check the possibility of the ontology to be used as key term sources for assessment questions. The results in Section 3.3.3 reveal that most of the concepts matched with assessment question keywords using n-gram string similarity measure and DCS. However, the results also provide evidence that some ontology concepts were not represented using proper naming conventions. This suggests that improving the representation of ontology elements is another issue requiring further consideration. We have suggested a guideline for proper concept naming in ontology especially course ontology for question generation. This work also contributes to the validating of the ontology concept completeness using a textbook revision question as a competency question. This approached could identify whether the selected course ontology covers all concepts teaches in a particular subject. This is done thru the matching process between keywords in assessment question with concept exist in course ontology.

8.3.2 Contribution to novel QG strategies for generating assessment questions with categorized question templates.

The proposed question generation strategy to map categorized question templates with ontology element has contributed to the extension of automatic question generation strategies using ontology. Many research studies have explored the ontological question generation approach to finding distractors for MCQ, FIB questions, and verification question. Thus the question generation strategies proposed in this research for factual S/LAQ may provide extension features to existing automatic question generator. The proposed question generation strategies could be tested on more subjects in other domains, and upon the success of its implementation, the tools may be added as a plug-in for Protégé Ontology editor.

The design of categorised question templates discussed in Section 4.1 provides the possibility of adding more question templates based on the question categories, so as long as the question has the same structures as discussed in Section 5.2.1. Adding question templates may increase the number of questions generated and indirectly support the generation of dynamic questions as suggested by Lin and Chou, (2011) to reduce plagiarism among students.

8.3.3 Contribution to automatic QG architecture and framework

The architecture and framework for AQGen was proposed and discussed in Chapter 5, and shown to contribute in terms of using it as a guideline to design ontology-based question generator which integrates the pre-existing question templates and question strategies to generate questions that assess students at different question taxonomy. The AQGen framework was designed to answer RQ2 regarding how to generate question from course ontology. Rule-based techniques in NLP were used to develop each question generation modules as detailed in Section 5.2.3. In addition, the AQGen framework may be used as a tool to validate other course ontologies for the purpose of automatic question generator.

8.4 Limitations and Recommendation for Future Work

There are limitations in conducting experiments in this research that might have had an impact on the quality of research findings. The key limitations of this research are:

- **The availability of online course ontology**: to generalise the view that the proposed question generation strategies might able to generate acceptable questions, the techniques needed to be evaluated for a variety of course ontology. Most domain ontologies are not designed for educational syllabi. Some course ontologies identified, were created for specific experiment purposes and not available for public access and viewing. A request for course ontology was made to an Object-Oriented Programming subject. Unfortunately, he was not able to share the ontology file.
- **The number of participants:** A questionnaire was designed to obtain feedback from experts on the acceptability measure of generated questions. Volunteer participants were sought; however only 17 experts responded, 19 were not willing to be involved, and 14 did not return the questionnaire.
- **Feedback and solution:** Feedback and solution to the questions are regarded by the educator as an essential element that influenced learning. Although learning through answering repetitive and

dynamic questions model proposed in this work could significantly improve learning, the research also concluded significance success of using feedback and solution model to develop student learning and motivation (Ammons, 1956; Kulhavy & Wager, 1993). Therefore, by adding feedback and solution elements into the existing system would widen the use of the system to cover different learning style.

These limitations raise a few challenges with regard to further experiments in the future. Recommendations for future work are listed below:

- The assessment questions collected to analyse the type of questions for generated assessment questions were based on one subject, that of Operating Systems. The questions were taken from OS textbook's reviewed questions and five past year examination papers from a university bachelor's degree programs. Therefore, the question is limited to three categories which are definition, concept completion and comparison. Analysing other question categories would widen the selection of question categories for assessing students.
- The AQGen framework applied and evaluated two computer science subjects. More questions from different subjects should be analysed to evaluate the generality of the proposed strategies. However, choosing the right subjects that only assess the factual knowledge should be considered.
- The results for acceptability measure on AQGen generated questions indicate a need for improvement. The improvement that needs to be explored includes considering other question templates, extending the simple question to complex questions, and exploring difference question structures.
- The proposed topic-based question selection was applied only to show the feasibility of the AQGen tool for the end-user. It is not the best question selection algorithm, but the purpose is to filter questions and avoid over-generation. Exploring better question selection algorithms would maximise the benefits of AQGen.

- The future experiment should compare the effect of learning through feedback and solution with learning by using repetitive and dynamic questions. The comparison could examine the independent variable such 1) How the two model of learning change learning strategy, 2) Which of these two models improve motivation for learning and finally 3) Which of these two models could improve overall performance for learning.?

REFERENCES

AFZAL, N. & MITKOV, R., 2014. Automatic generation of multiple choice questions using dependency-based semantic relations. *Soft Computing,* 18**,** 1269-1281.

AFZAL, N. & PEKAR, V. 2009. Unsupervised relation extraction for automatic generation of multiple-choice questions. Proceedings of the International Conference RANLP-2009, 1-5.

AL-YAHYA, M. 2011. OntoQue: a question generation engine for educational assessment based on domain ontologies. 11th IEEE International Conference on Advanced Learning Technologies (ICALT). IEEE, 393-395.

AL-YAHYA, M., 2014. Ontology-Based Multiple Choice Question Generation. *The Scientific World Journal,* 2014**,** 9 pages.

AL-YAHYA, M., AL-KHALIFA, H., BAHANSHAL, A., AL-ODAH, I. & AL-HELWAH, N. 2010. An ontological model for representing semantic lexicons: an application on time nouns in the holy Quran. *Arabian Journal for Science and Engineering,* 35**,** 21.

AL-YAHYA, M., GEORGE, R. & ALFARIES, A., 2015. Ontologies in E-learning: a review of the literature. *International Journal of Software Engineering and Its Applications,* 9**,** 67-84.

ALI, H., CHALI, Y. & HASAN, S. A., 2010. Automation of question generation from sentences. Proceedings of QG2010: The Third Workshop on Question Generation,. 58-67.

ALSUBAIT, T., PARSIA, B. & SATTLER, U., 2014. Generating multiple choice questions from ontologies: How Far Can We Go? International Conference on Knowledge Engineering and Knowledge Management. Springer, 66-79.

ALSUBAIT, T., PARSIA, B. & SATTLER, U., 2013. A similarity-based theory of controlling MCQ difficulty. In Second International

Conference on e-Learning and e-Technologies in Education (ICEEE), pp. 283-288. IEEE

AMMONS, R.B., 1956. Effects of knowledge of performance: A survey and tentative theoretical formulation. *The Journal of general psychology*, *54*(2), pp.279-299.

ANDERSEN, U., BRAASCH, A., HENRIKSEN, L., HUSZKA, C., JOHANNSEN, A., KAYSER, L., MAEGAARD, B., NORGAARD, O., SCHULZ, S. & WEDEKIND, J., 2012. Creation and Use of Language Resources in a Question-Answering eHealth System. LREC. 2536-2542.

ANDERSON, L. W. & SOSNIAK, L. A., 1994. *Bloom's taxonomy*, Univ. Chicago Press.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S. & EPPIG, J. T. 2000. Gene Ontology: tool for the unification of biology. *Nature genetics,* 25**,** 25.

BACH, T.-L. & DIENG-KUNTZ, R., 2005. Measuring Similarity of Elements in OWL DL ontologies. Workshop on Contexts and Ontologies: Theory, Practice and Applications.

BALCH, W. R., 2007. Effects of test expectation on multiple-choice performance and subjective ratings. *Teaching of Psychology,* 34(4), pp. 219-225.

BECHHOFER, S., HORROCKS, I., GOBLE, C. & STEVENS, R., 2001. OilEd: a reason-able ontology editor for the semantic web. Annual Conference on Artificial Intelligence. Springer, 396-408.

BECHHOFER, S., HORROCKS, I. & PATEL-SCHNEIDER, P. 2003. People & Pet. https://protegewiki.stanford.edu/wiki/Protege_ Ontology _Library. Retrieved on January 2016.

BELL, B. & COWIE, B. 2001. The characteristics of formative assessment in science education. *Science education,* 85**,** 536-553.

BENNETT, R. E. 2011. Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice,* 18**,** 5-25.

BERGMAN, M.K., 2010. A Reference Guide to ontology best practice. AI3 Adaptive Information blog, September 13, 2010

BERNERS-LEE, T., HENDLER, J. & LASSILA, O., 2001. The semantic web. *Scientific American,* 284**,** 34-43.

BIGGS, J., 2002. Aligning Teaching and Assessment to Curriculum Objectives. Imaginative Curriculum Guide. Available: http://www.itslifejimbutnotasweknowit.org.uk/files/CPLHE/Biggs%20Constructive%20alignment.rtf. Retrieved on February 2016

BLOOM, B. S., 1956. Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I: Cognitive Domain. New York, US: D. Mckay.

BODOFF, D. & RABAN, D., 2016. Question types and intermediary elicitations. *Journal of the Association for Information Science and Technology,* 67**,** 289-304.

BREWSTER, C. & O'HARA, K., 2007. Knowledge representation with ontologies: Present challenges—Future possibilities. *International Journal of Human-Computer Studies,* 65**,** 563-568.

BRICKLEY, D. & MILLER, L., 2010. FOAF vocabulary specification 0.91. Citeseer.

BROADFOOT, P. & BLACK, P., 2004. Redefining Assessment? The First Ten Years of Assessment in Education. *Assessment in Education: Principles, Policy & Practice,* 11**,** 7-26.

BROWN, J. C., FRISHKOFF, G. A. & ESKENAZI, M., 2005. Automatic question generation for vocabulary assessment. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 819-826.

BUITELAAR, P., EIGNER, T. & DECLERCK, T., 2004. OntoSelect: A dynamic ontology library with support for ontology selection. In Proceedings of the Demo Session at the International Semantic Web Conference. Citeseer.

CARCHIOLO, V., LONGHEU, A. & MALGERI, M., 2010. Reliable peers and useful resources: Searching for the best personalised learning path in a trust-and recommendation-aware environment. *Information Sciences,* 180**,** 1893-1907.

CASTELLANOS-NIEVES, D., FERNÁNDEZ-BREIS, J. T., VALENCIA-GARCÍA, R., MARTÍNEZ-BÉJAR, R. & INIESTA-MORENO, M. 2011. Semantic web technologies for supporting learning assessment. *Information Sciences,* 181**,** 1517-1537.

CHALI, Y. & HASAN, S. A., 2015. Towards topic-to-question generation. *Computational Linguistics,* 41**,** 1-20.

CHI, M. T. H., DE LEEUW, N., CHIU, M.-H. & LAVANCHER, C. 1994. Eliciting self-explanations improves understanding. *Cognitive science,* 18**,** 439-477.

CHIN, C. 2006. Using self-questioning to promote pupils' process skills thinking. *School Science Review,* 87**,** 113-119.

CHIN, C. & BROWN, D. E., 2000. Learning deeply in science: An analysis and reintegration of deep approaches in two case studies of grade 8 students. *Research in Science Education,* 30**,** 173-197.

CHIN, C. & KAYALVIZHI, G., 2005. What do pupils think of open science investigations? A study of Singaporean primary 6 pupils. *Educational Research,* 47**,** 107-126.

COHEN, F. S., 1929. What is a Question? *The monist,* 39**,** 350-364.

COHEN, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement,* 20**,** 37-46.

COHEN, J., 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin,* 70**,** 213.

CONSORTIUM, G. O., 2016. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic acids research,* 45**,** D331-D338.

CONSORTIUM, P. O., 2002. The Plant Ontology™ consortium and plant ontologies. *Comparative and Functional Genomics,* 3**,** 137-142.

CUBRIC, M. & TOSIC, M., 2011. Towards automatic generation of e-assessment using semantic web technologies. *International Journal of e-Assessment,* 1.

CUNNNINGHAM, G., 2005. Assessment in the classroom: constructing and interpreting texts. Routledge

DIXSON, D. D. & WORRELL, F. C., 2016. Formative and summative assessment in the classroom. *Theory into practice,* 55**,** 153-159.

DOUGIAMAS, M. & TAYLOR, P., 2003. Moodle: Using Learning Communities to Create an Open Source Course Management System. EdMedia: World Conference on Educational Media and Technology. *Association for the Advancement of Computing in Education (AACE)*, 171-178.

DOUGIAMAS, M. & TAYLOR, P. C., 2002. Interpretive Analysis of an Internet-based Course Constructed using a New Courseware Tool Called Moodle. 2nd Conference of HERDSA (The Higher Education Research and Development Society of Australasia). 7-10.

ERDOGAN, I., 2017. Turkish Elementary Students' Classroom Discourse: Effects of Structured and Guided Inquiry Experiences that Stimulate Student Questions and Curiosity. *International Journal of Environmental and Science Education,* 12(5), pp. 1111-1137.

FAIZAN, A. & LOHMANN, S., 2018. Automatic Generation of Multiple Choice Questions from Slide Content using Linked Data. *In Procedings of the 8th International Conference on Web Intelligence, Mining and Semantic,* ACM, p. 32.

FELLBAUM, C., 1998. *WordNet : An Electronic Lexical Database,* Cambridge, MIT Press.

FENSHAM, P. J., GUNSTONE, R. F. & WHITE, R. T., 2013. *The content of science: A constructivist approach to its teaching and learning*, Routledge.

FERNÁNDEZ-BREIS, J. T. & MARTIINEZ-BEJAR, R., 2002. A cooperative framework for integrating ontologies. *International Journal of Human-Computer Studies,* 56**,** 665-720.

FLEISS, J. L., 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin,* 76**,** 378-382.

FUNK, S. C. & DICKSON, K. L., 2011. Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology,* 38(4), pp. 273-277.

GRAESSER, A. C., D'MELLO, S., HU, X., CAI, Z., OLNEY, A. & MORGAN, B. 2012. AutoTutor. *Applied natural language processing: Identification, investigation and resolution.* IGI Global.

GRAESSER, A. C., JACKSON, G. T., MATTHEWS, E. C., MITCHELL, H. H., OLNEY, A., VENTURA, M., CHIPMAN, P., FRANCESCHETTI, D., HU, X. & LOUWERSE, M. M., 2003. Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog. Proceedings of the Annual Meeting of the Cognitive Science Society.

GRAESSER, A. C., LU, S., JACKSON, G. T., MITCHELL, H. H., VENTURA, M., OLNEY, A. & LOUWERSE, M. M., 2004. AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers,* 36**,** 180-192.

GRAESSER, A. C. & OLDE, B. A., 2003. How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology,* 95**,** 524.

GRAESSER, A. C. & PERSON, N. K. 1994. Question Asking During Tutoring. *American Educational Research Journal,* 31**,** 104-137.

GRAESSER, A. C., PERSON, N. K. & HUBER, J. D. 1992. Mechanisms that generate questions. En T. Lauer, E. Peacock, & AC Graesser (Eds.), Questions and Information Systems (pp. 167±187). Hillsdale, NJ: Erlbaum.

GRUBER, T. R., 1993. A translation approach to portable ontology specifications. *Knowledge acquisition,* 5**,** 199-220.

GRUBER, T. R. 1995. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies,* 43**,** 907-928.

GRÜNINGER, M. & FOX, M. S., 1995. The role of competency questions in enterprise engineering. *Rolstadås A. (eds) Benchmarking — Theory and Practice. IFIP Advances in Information and Communication Technology.* Boston, MA: Springer.

GWET, K. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters,* Gaithersburg, MD, USA, Advanced Analytics, LLC.

GWET, K. L., 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology,* 61**,** 29-48.

HALPERN, D. F., MILLIS, K., GRAESSER, A. C., BUTLER, H., FORSYTH, C. & CAI, Z., 2012. Operation ARA: A computerized learning game that teaches critical thinking and scientific reasoning. *Thinking Skills and Creativity,* 7**,** 93-100.

HEILMAN, M., 2011. Automatic factual question generation from text. *Language Technologies Institute School of Computer Science Carnegie Mellon University,* 195.

HEILMAN, M. & SMITH, N. A., 2009. Question generation via overgenerating transformations and ranking. CARNEGIE-MELLON UNIV PITTSBURGH PA LANGUAGE TECHNOLOGIES INST.

HENDLER, J. & MCGUINNESS, D. L., 2000. The DARPA agent markup language. *IEEE Intelligent systems,* 15**,** 67-73.

HOLOHAN, E., MELIA, M., MCMULLEN, D. & PAHL, C. 2005. Adaptive e-learning content generation based on semantic web technology.

HOLOHAN, E., MELIA, M., MCMULLEN, D. & PAHL, C., 2006. The generation of e-learning exercise problems from subject ontologies. *Sixth International Conference on Advanced Learning Technologies. IEEE,* 967-969.

HOSHINO, A. & NAKAGAWA, H., 2005. A real-time multiple-choice question generation for language testing: a preliminary study. Proceedings of the second workshop on Building Educational Applications Using NLP. Association for Computational Linguistics, 17-20.

HOVY, E., GERBER, L., HERMJAKOB, U., LIN, C.-Y. & RAVICHANDRAN, D., 2001. Toward semantics-based answer pinpointing. Proceedings of the First International Conference on Human Language Technology Research. Association for Computational Linguistics, 1-7.

HU, X. & GRAESSER, A. C. 2004. Human use regulatory affairs advisor (HURAA): Learning about research ethics with intelligent learning modules. *Behavior Research Methods, Instruments, & Computers,* 36**,** 241-249.

IBRAHIM TEO, N. H. & JOY, M., 2016. Evaluation of an automatic question generation approach using ontologies. In European Conference on e-Learning p.735. Academic Conferences International Limited.

KEARSLEY, G. P., 1976. Questions and question asking in verbal discourse: A cross-disciplinary review. *Journal of psycholinguistic research,* 5**,** 355-375.

KLYNE, G. C. & MCBRIDE, J. B., 2003. Resource Description Framework (RDF): Concepts and Abstract Syntax. *IN MCBRIDE, B.(Ed.).*

KNUBLAUCH, H., FERGERSON, R. W., NOY, N. F. & MUSEN, M. A. The Protégé OWL plugin: An open development environment for semantic web applications. International Semantic Web Conference, 2004. Springer, 229-243.

KULHAVY, R.W. and WAGER, W., 1993. Feedback in programmed instruction: Historical context and implications for practice. *Interactive instruction and feedback, 320*.

KUMAR, A. N., 2005. Generation of problems, answers, grade, and feedback case study of a fully automated tutor. *Journal on Educational Resources in Computing (JERIC),* 5**,** 3.

LE, N.-T., NGUYEN, N.-P., SETA, K. & PINKWART, N. 2014. Automatic question generation for supporting argumentation. *Vietnam Journal of Computer Science,* 1**,** 117-127.

LEO, J., KURDI, G., MATENTZOGLU, N., PARSIA, B., SATTLER, U., FORGE, S., DONATO, G. and DOWLING, W., 2019. Ontology-Based Generation of Medical, Multi-term MCQs. *International Journal of Artificial Intelligence in Education*, *29*(2), pp.145-188

LEVY, R. & ANDREW, G., 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. *Proceedings of the Fifth International Conference on Language Resources and Evaluation. Citeseer,* 2231-2234.

LIN, C.-Y., 2008. Automatic question generation from queries. *Workshop on the Question Generation Shared Task.* 156-164.

LIN, C.-Y. & CHOU, C.-C. 2011. Effectiveness of Ontology-Based Online Exam Platform for Programming Language Education. *Computing in Civil Engineering.*

LIU, C.-L., WANG, C.-H., GAO, Z.-M. & HUANG, S.-M., 2005. Applications of lexical information for algorithmically composing multiple-choice cloze items. *Proceedings of the Second Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, 1-8.

MA, Y., GNANESWARAN, S., HARDAS, M. & I.KHAN, J. 2006. Ontology of Operating System. *Internetworking and Media Communications Research Laboratories.* Internetworking and Media Communications Research Laboratories , Department of Computer Science: Kent State University.

MILLIS, K., FORSYTH, C., BUTLER, H., WALLACE, P., GRAESSER, A. & HALPERN, D. 2011. Operation ARIES!: A serious game for teaching scientific inquiry. *Serious games and edutainment applications.* Springer.

MITKOV, R., AN HA, L. & KARAMANIS, N. 2006. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering,* 12**,** 177-194.

MURUGAN, S. R., BALA, R. P. & AGHILA, G. 2013. An ontology for exploring knowledge in computer networks. *International Journal on Computational Sciences & Applications (IJCSA),* 3**,** 13-21.

MUSEN, M. A. 2015. The Protege project: A look back and a look forward. *AI Matters: Association of Computing Machinery Specific Interest Group in Artificial Intelligence,* 1**,** 4-12.

NAZ, M., ASLAM, M. & DAR, E. 2010. Utility based agent for test paper generation. *International Journal of Multidisciplinary Sciences and Engineering,* 1**,** 1-5.

NICOL, D. J. & MACFARLANE-DICK, D. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education,* 31**,** 199-218.

NOY, N. F. & MCGUINNESS, D. L. 2001. Ontology development 101: A guide to creating your first ontology. *Stanford Knowledge Systems Laboratory Technical report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880*, Stanford, CA.

NOY, N. F., SHAH, N. H., WHETZEL, P. L., DAI, B., DORF, M., GRIFFITH, N., JONQUET, C., RUBIN, D. L., STOREY, M.-A. & CHUTE, C. G. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research,* 37**,** W170-W173.

OLNEY, A. M., GRAESSER, A. C. & PERSON, N. K. 2012. Question generation from concept maps. *Dialogue & Discourse,* 3**,** 75-99.

OTERO, J. & GRAESSER, A. C. 2001. PREG: Elements of a model of question asking. *Cognition and instruction,* 19**,** 143-175.

PAPASALOUROS, A., KANARIS, K. & KOTIS, K., 2008. Automatic Generation Of Multiple Choice Questions From Domain Ontologies. e-Learning. Citeseer, 427-434.

PEDROSA DE JESUS, H., NERI DE SOUZA, F., TEIXEIRA-DIAS, J. J. C. & WATTS, M. 2005. Organising the chemistry of question-based learning: A case study. *Research in Science & Technological Education,* 23**,** 179-193.

PINCKARD, R. N., MCMAHAN, C.A., PRIHODA, T.J., LITTLEFIELD, J.H., & JONES, A.C., 2009. Short-answer examinations improve student performance in an oral and maxillofacial pathology course. *Journal of Dental Education,* 73(8), pp. 950-961.

PIWEK, P. & BOYER, K. E., 2012. Varieties of question generation: introduction to this special issue. *Dialogue & Discourse,* 3**,** 1-9.

RICHARDS, D., 2009. A social software/Web 2.0 approach to collaborative knowledge engineering. *Information Sciences,* 179**,** 2515-2523.

RUS, V., CAI, Z. & GRAESSER, A. 2008. Question generation: Example of a multi-year evaluation campaign. *Proc WS on the QGSTEC*.

RUS, V., D'MELLO, S., HU, X. & GRAESSER, A. 2013. Recent advances in conversational intelligent tutoring systems. *AI magazine,* 34**,** 42-54.

SADLER, D. R., 1998. Formative assessment: Revisiting the territory. *Assessment in education: principles, policy & practice,* 5**,** 77-84.

SILBERSCHATZ, A., GAGNE, G. & GALVIN, P. B., 2004. *Operating System Concepts with Java*, John Wiley & Sons.

SILBERSCHATZ, A., GAGNE, G. & GALVIN, P. B., 2012. *Operating System Concepts,* New Jersey, John Wiley & Sons, Inc.

SILBERSCHATZ, A., GALVIN, P. B. & GAGNE, G. 2002. *Operating System Concepts,* Wiley.

SINGER, H. & DONLAN, D. 1982. Active comprehension: Problem-solving schema with question generation for comprehension of complex short stories. *Reading Research Quarterly***,** 166-186.

SMITH, B., ASHBURNER, M., ROSSE, C., BARD, J., BUG, W., CEUSTERS, W., GOLDBERG, L. J., EILBECK, K., IRELAND, A. & MUNGALL, C. J. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology,* 25**,** 1251.

SOE, T. L. 2014. Ontology-Based Indexing and Semantic Indexing in Information Retrieval Systems. *International Journal of Research Studies in Computer Science and Engineering (IJRSCSE),* 1**,** 1-9.

SOLDATOVA, L. N. & MIZOGUCHI, R. 2009. An Ontology-Based Test Generation System. *In:* DICHEVA, D., MIZHOGUCCI, R. & GREER, J. (eds.) *Semantic Web Technologies for E-learning.* Amsterdam, Netherland: IOS Press.

STAHL, S. A. 1986. Three principles of effective vocabulary instruction. *Journal of Reading,* 29**,** 662-668.

STIGGINS, R. J., CONKLIN, N. F. & BRIDGEFORD, N. J. 1986. Classroom assessment: A key to effective education. *Educational Measurement: issues and practice,* 5**,** 5-17.

SULTAN, A. B. M. 2010. Shuffling algorithms for automatic generator question paper system. *Computer and Information Science,* 3**,** 244.

SURE, Y., ANGELE, J. & STAAB, S., 2002. OntoEdit: Guiding ontology development by methodology and inferencing. OTM Confederated *International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 1205-1222.

TANENBAUM, A. S. 2001. *Modern Operating Systems*, Prentice Hall.

TEO, N. H. I., BAKAR, N. A. & KARIM, S., 2012. Designing GA-based auto-generator of examination questions. Sixth UKSim/AMSS European Symposium on Computer Modeling and Simulation (EMS). IEEE, 60-64.

TOSIC, M. & CUBRIC, M., 2009. SeMCQ–Protégé plugin for automatic ontology-driven multiple choice question tests generation. Procs of the 11th International Protege Conference. Stanford Center for Biomedical Informatics Research.

TSURUOKA, Y. 2006. GENIA tagger: Part-of-speech tagging, shallow parsing, and named entity recognition for biomedical text. *Available at: www-tsujii. is. su-tokyo. ac. jp/GENIA/tagger*.

VANLEHN, K., GRAESSER, A. C., JACKSON, G. T., JORDAN, P., OLNEY, A. & ROSÉ, C. P. 2007. When are tutorial dialogues more effective than reading? *Cognitive science,* 31**,** 3-62.

VILJANEN, K., TUOMINEN, J. & HYVÖNEN, E., 2009. Ontology libraries for production use: The Finnish ontology library service ONKI. European Semantic Web Conference. Springer, 781-795.

VINU, E. V. & KUMAR, S., 2015. A novel approach to generate MCQs from domain ontology: Considering DL semantics and open-world assumption. *Web Semantics: Science, Services and Agents on the World Wide Web,* Volume 34, pp. 40-54.

WATTS, M., ALSOP, S., GOULD, G. & WALSH, A. 1997. Prompting teachers' constructive reflection: Pupils' questions as critical incidents. *International Journal of Science Education,* 19**,** 1025-1037.

WOLFE, J. H. 1976. Automatic question generation from text-an aid to independent study. *ACM SIGCUE Outlook.* ACM.

WONGPAKARAN, N., WONGPAKARAN, T., WEDDING, D. & GWET, K. L. 2013. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research,* 13**,** 61.

WOO, S., LI, Z. & MIRKOVIC, J., 2016. Good automatic authentication question generation. Proceedings of the 9th International Natural Language Generation conference. 203-206.

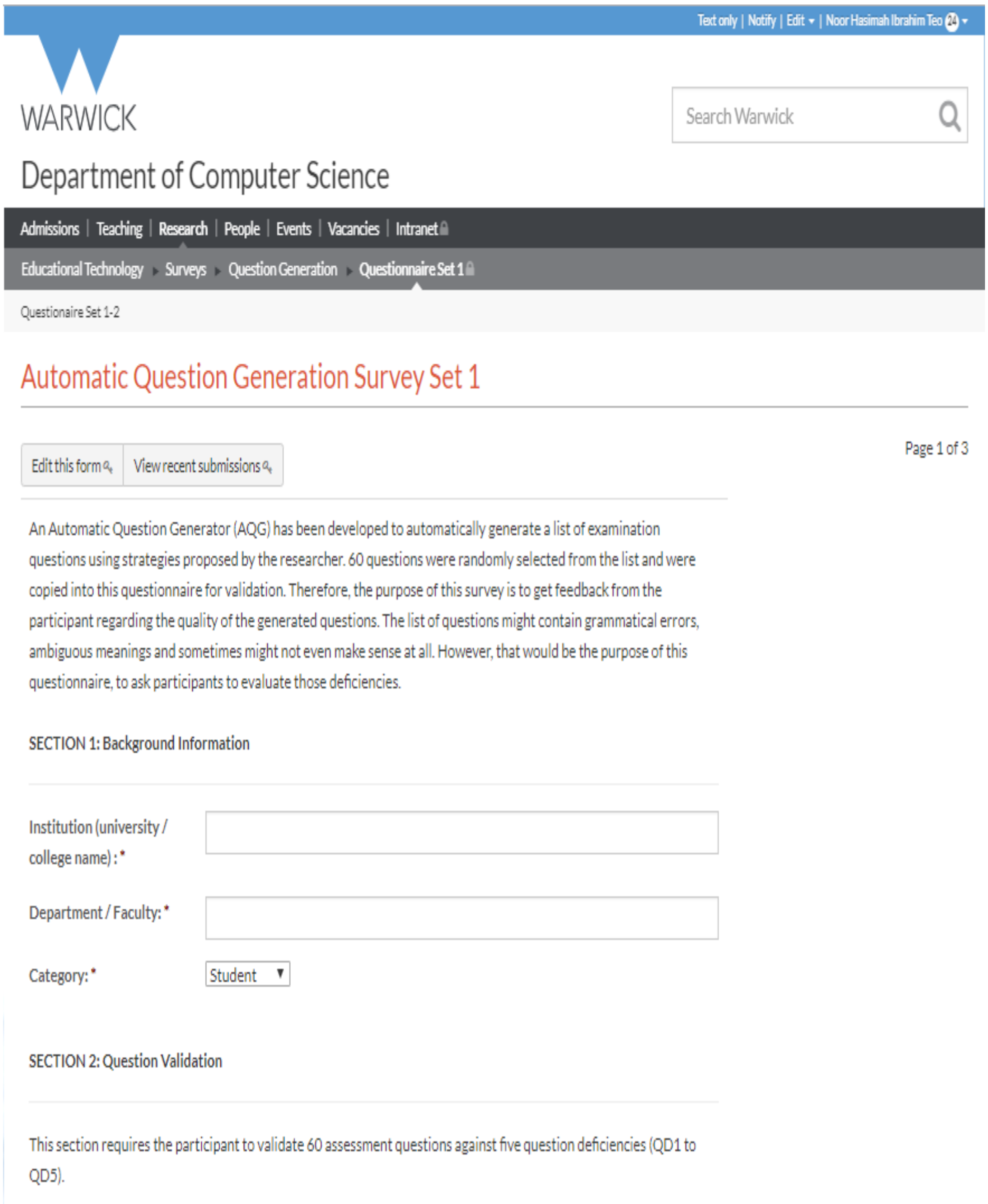
XIONG, L. & SHI, J., 2010. Automatic generating test paper system based on genetic algorithm. Second International Workshop on Education Technology and Computer Science. 272-275.IEEE

YORKE, M., 2003. Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher education,* 45**,** 477-501.

ZHENG, Z., SI, X., CHANG, E. & ZHU, X. K., 2011. Generating natural language questions from keywords with user refinements. *Proceedings of 5th International Joint Conference on Natural Language Processing.* 947-955.

Appendix A

Questionnaire Survey on the Acceptability Measure for Automated Generated Questions



Text only | Notify | Edit | Noor Hasimah Ibrahim Teo

WARWICK

Search Warwick

Department of Computer Science

Admissions | Teaching | Research | People | Events | Vacancies | Intranet

Educational Technology › Surveys › Question Generation › Questionnaire Set 1

Questionaire Set 1-2

Automatic Question Generation Survey Set 1

Page 1 of 3

Edit this form | View recent submissions

An Automatic Question Generator (AQG) has been developed to automatically generate a list of examination questions using strategies proposed by the researcher. 60 questions were randomly selected from the list and were copied into this questionnaire for validation. Therefore, the purpose of this survey is to get feedback from the participant regarding the quality of the generated questions. The list of questions might contain grammatical errors, ambiguous meanings and sometimes might not even make sense at all. However, that would be the purpose of this questionnaire, to ask participants to evaluate those deficiencies.

SECTION 1: Background Information

Institution (university / college name) : *

Department / Faculty: *

Category: * Student

SECTION 2: Question Validation

This section requires the participant to validate 60 assessment questions against five question deficiencies (QD1 to QD5).

Text only | Notify | Edit | Noor Hasimah Ibrahim Teo

WARWICK

Search Warwick

Department of Computer Science

Admissions | Teaching | Research | People | Events | Vacancies | Intranet

Educational Technology ▸ Surveys ▸ Question Generation ▸ Question Validation

Question Deficiencies Description

Please read the following description of QDs to help you understand the meaning.

Question deficiencies(QD)	**Descriptions**
QD1: This question is grammatically correct.	If you believe the question is grammatically correct, please select **YES**. If you are not sure, please select **NOT SURE**. If you believe the question has a grammatical error(s), please select **NO** and state the reason.
QD2: This question is understandable.	If you understand the question, please select **YES**. If you are uncertain about the question, please select **NOT SURE**. If you do not understand due to the technical terms used, please select **NO** and state the reason.
QD3: This question is within the context.	If you believe the question is within the context of this subject, please select **YES**. If you are not sure, please select **NOT SURE**. If the question is not appropriate for this subject, please select **NO** and state the reason.
QD4: This question is useful.	If you believe the question is helpful in assessing student understanding, please select **YES**. If you are not sure, please select **NOT SURE**. If you believe the question is not helpful for assessment purposes, please select **NO** and state the reason.
QD5: The question word is accurate.	**Question word** is a function word used to ask a question, such as '*what*', '*explain*', and '*define*'. If you believe the question word is correct, please select **YES**. If you are not sure, please select **NOT SURE**. If you believe the question use wrong question word, please select **NO** and state the reason.

Department of Computer Science

Admissions | Teaching | Research | People | Events | Vacancies | Intranet

Educational Technology ▸ Surveys ▸ Question Generation ▸ Questionnaire Set 1

Questionaire Set 1-2

Choose **ONLY ONE** option for each of the QD in a question.

1. Define Computer Network.

.*	YES	NOT SURE	NO
QD1: This question is grammatically correct.	○	○	○
QD2: This question is understandable.	○	○	○
QD3: This question is within the context.	○	○	○
QD4: This question is useful.	○	○	○
QD5: The question word is accurate.	○	○	○

If NO, please state reason(s)

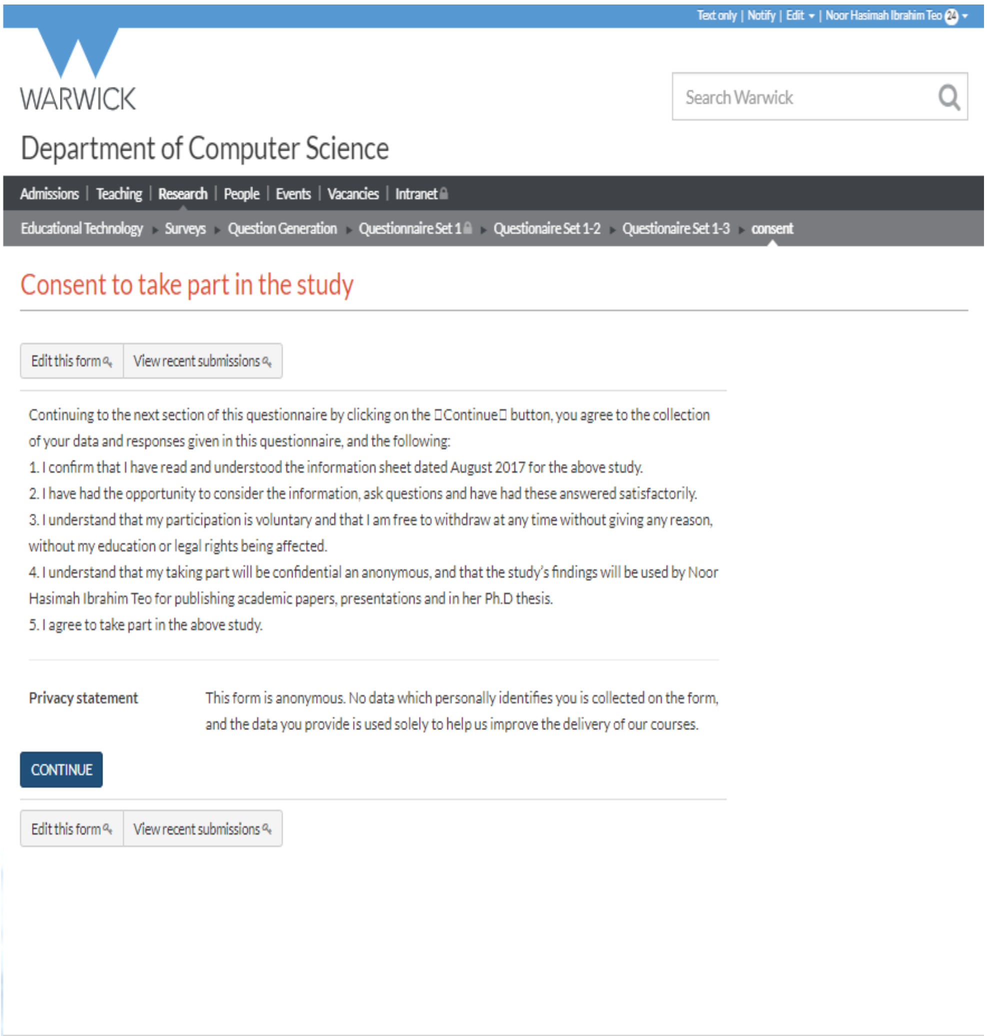
QD1:
QD2:
QD3:
QD4:
QD5:

2. Define switch.

.*	YES	NOT SURE	NO
QD1: This question is grammatically correct.	○	○	○
QD2: This question is understandable.	○	○	○
QD3: This question is within the context.	○	○	○
QD4: This question is useful.	○	○	○
QD5: The question word is accurate.	○	○	○

If NO, please state reason(s)

QD1:
QD2:
QD3:
QD4:
QD5:



Text only | Notify | Edit | Noor Hasimah Ibrahim Teo

WARWICK

Search Warwick

Department of Computer Science

Admissions | Teaching | Research | People | Events | Vacancies | Intranet

Educational Technology ▸ Surveys ▸ Question Generation ▸ Questionnaire Set 1 ▸ Questionaire Set 1-2 ▸ Questionaire Set 1-3 ▸ consent

Consent to take part in the study

Edit this form | View recent submissions

Continuing to the next section of this questionnaire by clicking on the Continue button, you agree to the collection of your data and responses given in this questionnaire, and the following:

1. I confirm that I have read and understood the information sheet dated August 2017 for the above study.
2. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.
3. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason, without my education or legal rights being affected.
4. I understand that my taking part will be confidential an anonymous, and that the study's findings will be used by Noor Hasimah Ibrahim Teo for publishing academic papers, presentations and in her Ph.D thesis.
5. I agree to take part in the above study.

Privacy statement This form is anonymous. No data which personally identifies you is collected on the form, and the data you provide is used solely to help us improve the delivery of our courses.

CONTINUE

Edit this form | View recent submissions

Appendix B

List of Generated Questions used in Questionnaire Survey

Appendix B1: Questions SET 1

Category	Questions
Definition	Define **computer network.** Define **switch**. Define **client server**. Define **Ethernet**. Define **internet protocol**.
	What is **communication media**? What is **scale**? What is **extranet**? What is **firewall**? What is **public key cryptography**?
	What does it means by **Network Operating System**? What does it means by **digital signature**? What does it means by **malicious software**? What does it means by **worm**? What does it means by **honeypot**?
	Define **tunnelling.** Define **star mesh**. Define **ssh-2**. Define **user control**. Define **blowfish**.
	What is **idea**? What is **wiretapping**? What is **backdoor**? What is **interruption**? What is **smurf attack**?
	What does it means by **svoip**? What does it means by **4-g**? What does it means by **bootp**? What does it means by **802-3-ab**? What does it means by **FM**?
Concept Completion	Explain the term **OSI model**. Explain the term **infrared**. Explain the term **twisted pair**. Explain the term **wlan**. Explain the term **bridge.**
	Discuss what a **protocol** is.

	Discuss what a **hub** is. Discuss what a **peer to peer** is. Discuss what a **presentation** is. Discuss what a **gigabit Ethernet** is.
	Explain the term **b-a-cnet**. Explain the term **dhcp**. Explain the term **telnet**. Explain the term **AM**. Explain the term **access control**.
	Discuss what a **storage area network** is. Discuss what a **service control** is. Discuss what a **rfc** is. Discuss what a **Trojan horse** is. Discuss what a **boot sector virus** is.
Comparison	*(no template)*
Category	
Concept Completion	What is a kind of **gigabit Ethernet**? What is a version of **Ethernet**? What *belong to* **a WLAN**? What is a standard of **Ethernet**? What is a type of **Network Operating System**? What is deployed in **a world wide web**? What is invented by **Alcatel Lucent**? What transmit **a data**? What uses **a FTP**? What is a layer of **TCP/IP**?

Appendix B2: Questions SET 2

Category	Questions
Definition	Define **system software.** Define **i/o system**. Define **principle of i/o system**. Define **assembler**. Define **compiler**
	What is **secondary storage structure**? What is **a device driver**? What is **interrupt handler**? What is **direct memory access**? What is **disk scheduling**?
	What does it means by **concurrency control**? What does it means by **deadlock**? What does it means by **memory management**? What does it means by **virtual memory**? What does it means by **protection**?
	Define **avoidance and prevention.** Define **file protection**. Define **file concept**. Define **access method**. Define **buffering**.
	What is **mutual exclusion requirement**? What is **recovery**? What is **swapping**? What is **paging**? What is **multithreading**?
	What does it means by **context switching**? What does it means by **access matrix**? What does it means by **authentication**? What does it means by **one time password**? What does it means by **encryption**?
Concept Completion	Explain the term **file system**. Explain the term **security**. Explain the term **protection**. Explain the term **clock**? Explain the term **Operating System**?
	Discuss what a **linux system** is. Discuss what a **memory management** is. Discuss what a **virtual memory** is. Discuss what a **system software**is. Discuss what a **protection and security** is.
	Explain the term **process state**. Explain the term **system treads**. Explain the term **encryption**. Explain the term **revocation of access right**. Explain the term **memory partitioning**.
	Discuss what a **kernel** is. Discuss what a **linker** is. Discuss what a **overlay** is. Discuss what a **disk scheduling** is. Discuss what a **multitasking** is.

Appendix B3: Questions SET 3

Category	Questions
Definition	What is a application layer in TCP/IP suite? What is a ieee-802-11 in wireless? What is a [X] in [SC]?
	Define a bridge in networking hardware? Define aa access switch in switch? Define a passive attack in security ?
	In the context of public key cryptography, what is digital signature? In the context of topology, what is bus? In the context of packet switching, what is ATM?
	What is meant by the RSA in a Public-key encryption? What is meant by the confidentiality in a security? What is meant by the firewall in an Extranet?
	What is a IP address in Internet layer? What is a shared fibre in fibre optic? What is a cat – 3 in unshielded twisted-pair?
	Define dynamic routing in router? Define network hub in hub ? Define an access control in data link ?
	In the context of network, what is routing? In the context of session, what is synchronization? In the context of ethernet, what is 100-basefx?
	What is meant by the term token ring in a gigabit-ethernet? What is meant by the term bootp in a transmission control protocol? What is meant by the term ssh-2 in a ssh?
Concept Completion	Within the context of topology, explain the term hybrid. Within the context of firewall, explain the term bastion-host. Within the context of topology, explain the term bus.
	Within the context of firewall, explain the term Extranet. Within the context of local area network, explain the term storage area network. Within the context of transmission control protocol, explain the term DHCP.
Comparison	Differentiate between internet layer and transport layer. Differentiate between worm and virus. Differentiate between intranet and World Wide Web.
	What is the difference between internet and extranet? What is the difference between Metropolitan Area Network and Personal Area Network? What is the difference between token ring and gigabit ethernet?
	How does ATM differ from Sonet? How does peer-to-peer differ from client/server? How does stream cipher differ from block cipher?
	Describe the differences between router and switch. Describe the differences between gateway and modem. Describe the differences between infrared and microwave.
	Compare between optical fibre and twisted pair. Compare between internet layer and transport layer. Compare between passive attacks and active attacks.

	With regards to network hardware, explain the difference between bridge and brouter. With regards to multiplexing, explain the difference between circuit mode and statistical multiplexing. With regards to public key cryptography, explain the difference between digital signature and public key encryption.
	Differentiate between DNS and DHCP. Differentiate between FTP and HTTP. Differentiate between media access control and medium access control.
	What is the difference between netbios and b-a-cnet? What is the difference between bandwidth and throughput? What is the difference between cmac and hmac?
	How does packet differ from routing? How does flow control differ from error control? How does distance vector routing differ from edge router?
	Describe the differences between Ethernet hub and multiport hub. Describe the differences between active gateway and bidirectional gateway. Describe the differences between cat-3 and cat-5.
	Compare between bonded twisted pair and loaded twisted pair. Compare between translation bridge and transparent bridge. Compare between multiport hub and network hub.
	With regards to data link, explain the difference between flow control and error control. With regards to router, explain the difference between dynamic routing and static routing. With regards to LAN, explain the difference between optical repeaters and radio repeaters.

Appendix B4: Questions SET 4

Category	Questions
Definition	What is a I/O manager in system software? What is a disk management in secondary storage structure? What is a assembler in I/O system?
	Define a process management in Operating System ? Define a job scheduling in Mainframe Operating System ? Define an interrupt in Computer System Operation?
	In the context of I/O structure, what is synchronous I/O structure? In the context of clustered system, what is asymmetric clustering? In the context of Computer System structure, what is hardware protection?
	What is meant by the networking in a system component? What is meant by the Wide Area Network in a Network structure? What is meant by the process management in a CPU Scheduling?
	What is a primary storage in secondary storage management? What is a buffering in interprocess communication? What is a memory mapping in virtual memory?
	Define a one-to-one in multithreading models ? Define a non-preemptive in priority scheduling? Define a conflict phase in real time scheduling ?
	In the context of concurrency control, what is semaphore? In the context of secondary storage structure, what is disk management? In the context of fundamental of Operating System, what is buffering?
	What is meant by the term deadlock in a concurrency control? What is meant by the term file protection in a file system? What is meant by the term kernel in a basic architecture?
Concept Completion	Within the context of process and thread management, explain the term context switching.. Within the context of protection and security, explain the term protection. Within the context of security, explain the term authentication.
	Within the context of protection, explain the term access matrix. Within the context of basic architecture, explain the term shell. Within the context of deadlock, explain the term recovery. (30)
Comparison	Differentiate between clock hardware and clock software. Differentiate between assembler and compiler. Differentiate between swapping and paging.
	What is the difference between multiprogramming and multitasking? What is the difference between spooling and buffering? What is the difference between file support and file protection?
	How does multiprogramming differ from multitasking? How does segmentation differ from swapping? How does allocation algorithm differ from page replacement algorithm?
	Describe the differences between demand paging and overlay. Describe the differences between process state and process management. Describe the differences between program thread and system thread.

	Compare between process state and process management. Compare between assembler and compiler. Compare between security and protection.
	With regards to deadlock, explain the difference between detection and recovery. With regards to Operating System, explain the difference between buffering and spoofing. With regards to memory management, explain the difference between allocation algorithm and page replacement algorithm.
	Differentiate between paging and segmentation. Differentiate between swapping and paging. Differentiate between spooling and buffering.
	What is the difference between system thread and one-time password? What is the difference between context switching and multithreading? What is the difference between overlay and demand paging?
	How does unix/linux system differ from kernel? How does time sharing differ from spooling? How does detection differ from recovery?
	Describe the differences between clock hardware and clock software. Describe the differences between device controllers and direct memory access. Describe the differences between disk management and disk reliability.
	Compare between buffering and spooling. Compare between encryption and authentication. Compare between thread monitoring and system thread.
	With regards to memory management, explain the difference between paging and segmentation. With regards to concurrency control, explain the difference between deadlock and semaphore. With regards to I/O systems, explain the difference between assembler and compiler.