Towards Automatic Evaluation of Questions Generated from Ontologies

Samah Alkhuzaey[1,*,†], Floriana Grasso[1,†], Terry R. Payne[1,†] and Valentina Tamma[1,†]

[1]*University of Liverpool, L69 3BX, Liverpool, UK*

Abstract

Automatic question generation has emerged as an important field in educational technology. It enables the creation of large question banks for various learning environments. Nevertheless, the predominant reliance on human assessments to evaluate these generated questions hampers scalability and efficiency. To address this challenge, this paper presents an automatic framework that utilises ontological metrics to assess the complexity of questions generated from domain ontologies. The proposed approach is evaluated through an expert-based evaluation. The results reveal a consensus between the complexity scores generated by the framework and the opinions of educational experts, demonstrating the effectiveness of our proposed approach. However, the findings also highlight the need for adjustments to account for certain features that could enhance the accuracy of the proposed model's ratings.

Keywords

Question generation, ontology, evaluation, complexity

1. Introduction

In the field of artificial intelligence and educational technologies, there has been a focus on the development of systems capable of generating questions autonomously. This has led to the emergence of *Automatic Question Generation (AQG)* techniques, that address the challenges faced by examination question developers when creating a large number of educational questions. AQG utilises various sources of knowledge, both structured and unstructured, including text and semantic models such as ontologies and knowledge bases. Ontology-based AQG leverages semantic knowledge representations, or *ontologies* [1], to generate assessment questions with different characteristics. This use of ontologies as a semantic source offers several advantages over other generation models. Question generation models that utilise ontologies as a knowledge source have been found to exhibit better generalisation across domains and question formats than other generation models, such as those that rely on machine learning [2]. Furthermore, a crucial distinction that distinguishes ontology-based approaches from other AQGs is that the essential characteristics of a question, such as its difficulty, are not influenced by the language used in the question, but rather they are determined solely by the graph structure and the

EvalLAC'24: Workshop on Automatic Evaluation of Learning and Assessment Content, July 08, 2024, Recife, Brazil

*Corresponding author.

†These authors contributed equally.

S.Alkhuzaey@Liverpool.ac.uk (S. Alkhuzaey); .Grasso@liverpool.ac.uk (F. Grasso); T.R.Payne@liverpool.ac.uk (T. R. Payne); V.Tamma@liverpool.ac.uk (V. Tamma)

0000-0001-8883-1172 (S. Alkhuzaey); 0000-0001-8419-6554 (F. Grasso); 0000-0002-0106-8731 (T. R. Payne); 0000-0002-1320-610X (V. Tamma)

relationship between entities. This distances the generation and evaluation process from the external structure of the question, allowing for more focus on its semantics [3]. Despite the potential of using Ontology-based AQG systems, the evaluation of such approaches has to date primarily relied on human judgment, including the use of expert reviewers [4, 5, 6], students' performance [7, 8], or crowd-sourced evaluations [9]. While human judgment provides valuable insights into the quality and relevance of generated questions, it is subjective, time-consuming, and resource-intensive. Additionally, the scalability of human-based evaluation is limited, which hinders the comprehensive assessment of AQG systems across diverse domains and datasets. By automating the evaluation process, researchers can overcome the limitations of human-based assessment and improve efficiency, making evaluation efforts more scalable. Additionally, automated evaluation frameworks offer the potential to provide objective, reproducible, and quantifiable metrics for accurately measuring the performance of Ontology-based AQG systems.

Developing automatic measures to evaluate questions generated from ontologies is practical due to the structured and standardised nature of ontologies and the inherent consistency in their hierarchical organisation and relationships. Ontologies are formal representations of knowledge within a specific domain [1], comprising concepts, relationships, and rules that define their interaction. The structured nature of ontologies, which typically includes a hierarchical organisation of concepts and well-defined relationships, facilitates a systematic approach to question generation. Furthermore, the uniformity in structure across various ontologies facilitates the development of generalised methods for both question generation and its evaluation. The uniformity in the structure becomes clear when examining that most approaches that employ ontologies to generate questions tend to share common features, such as leveraging basic hierarchical relationships or other semantic elements (e.g. object properties). This consistency further supports the development of consistent evaluation measures.

In this paper, we investigate the possibility of using an automatic evaluation framework as a proxy for expert user evaluations when assessing the complexity level of question generated from ontologies. The different aspects involved in the construction of ontology-based questions are evaluated to understand their effect of question complexity. Our approach exploits the hierarchical structure and standardised relationships inherent in ontologies to establish a consistent evaluation method. In Section 2, we review similar research on evaluation methodologies in the field of ontology-based automatic question generation, before presenting our proposed framework in Section 3, where the theoretical foundations and proposed metrics are detailed in practical terms. The evaluation methodology is presented in Section 4, followed by preliminary results of our study in Section 5, and the conclusions in Section 6.

2. Related Work

Although Ontology-based AQG provides an automated approach for creating questions, it still requires significant input from educational experts to evaluate the generated questions. As with most natural language generation tasks, human evaluation is considered the benchmark against which the outcome of the generation process is compared. To assess the quality of questions generated from knowledge sources such as ontologies, various methods, metrics, and techniques have been employed. When evaluating the questions, quality can be examined across different

dimensions, such as the question's structure [5, 10, 11], cognitive level [8], difficulty [4, 10], semantic ambiguity [12, 8], practical usefulness in an educational context [4, 5, 12] or overall acceptability by an expert [13]. In general, quality assessment in ontology-based AQG can be broadly categorised as those related to the *language* of the question, and those associated with the question's *cognitive level*. Human-based evaluation continues to be widely used for assessing the effectiveness of Ontology-based AQG systems in both language and cognitive evaluations. Expert reviewers, who possess domain knowledge or expertise in exam construction, offer valuable insights into the appropriateness of the generated questions. Students can also be recruited to provide practical evaluations, allowing for feedback from end-users and reflecting the usability and comprehensibility of the generated questions in real educational contexts.

Experts are typically hired to evaluate questions based on linguistic aspects, such as grammatical correctness, syntactic consistency, and fluency [5, 10, 11, 14]. Grammatical correctness involves assessing whether the questions adhere to grammar rules, ensuring that they are error-free and clear for learners. Another linguistic metric commonly evaluated is the syntactic consistency of the questions. This involves ensuring that the questions have consistent syntactic features, such as the *Part of Speech (POS)* used. This measure ensures that the questions have a uniform syntactic structure, as syntactic inconsistencies may confuse learners. In this type of evaluation, experts are typically presented with a set of generated questions and asked to rate their quality based on specific criteria using a categorical scale.

Human-centred evaluations have been conducted to assess cognitive-level metrics, such as *question difficulty*, *discrimination*, and *complexity* [5, 6, 7, 8]. *Difficulty* and *discrimination* are usually measured using standard statistical analysis of responses, employing pedagogical theories such as *Item Response Theory (IRT)* [15]. This involves administering a subset of the generated questions to students in real or mock exams, where the actual difficulty and discrimination are calculated and compared with predicted values. *Difficulty* may be estimated by domain experts that draw on their knowledge and experience in the field [5, 6]; whereas *complexity* measures the inherent complexity of a question based on its structure, the cognitive processes required to answer it, and the depth of understanding it demands [8, 16]. Unlike statistical difficulty and discrimination, which are heavily influenced by learners' backgrounds and knowledge levels, *complexity* provides an intrinsic measure of a question's potential to engage and challenge learners.

While human evaluation is commonly seen as the most reliable method for assessing the quality of generated questions, it is not always practical, especially for systems that generate a large volume of questions, due to the substantial amount of time and effort needed to manually evaluate each question. Thus it is necessary to employ automated or semi-automated evaluation methods to ensure efficient and timely assessment. The fact that the average number of expert evaluators involved in these studies is typically three can further exacerbate these scalability issues [2]. Automatic evaluation techniques have thus gained attention as a means of addressing these limitations of human-centric assessment. Such methods use computational algorithms to analyse the generated questions, taking into account factors such as language and cognitive level. Notable examples include metrics that quantify the similarity of generated questions to those created by humans [12, 17]. Alsubait et al. [5], for example, developed specific similarity measures for ontologies to assess the similarity of distractors in Multiple Choice Questions (MCQs). Their assumption is that having similar choices increases the cognitive level required

Table 1
Common question templates with their corresponding RDF patterns. Class names are represented by upper-case letters, while instances are indicated by lower-case letters. 'P' represents any property type, 'OP' stands for object property and 'DP' for Datatype property. Constraints appear in bold.

#	Question Type	RDF Pattern	Abstract Specification	Example Question
01	Definition	<X> <rdf:type> <owl:Class> <X> <rdfs:comment> <string>	Define <X>	*Define a Coastal region*
02	Class Assertion	<x> <rdf:type> <X>	Is <x> an <X>?	*Is Nice a coastal city?*
03	Property Assertion	<x> <P> <y>	Is <x> <P> <y>?	*Is Paris the capital of France?*
04	MCQ	<x> <rdf:type> <X> <y> <rdf:type> <Y> <X>,<Y> <subClassOf> <Z>	Which of these is <X>:	*Which of these is a Rural area: A: Lyon B: Auvergne*
05	Complex MCQ	<x> <P1> <y> <x> <P2> <z>	What <P1> <y> and <P2> <z>?	*Which city has a population of over one million and borders the Mediterranean Sea?*
06	Aggregate	<x> <OP> <y> <y> <rdf:type> <Y> **COUNT** <y>	How many <Y> does <x> have?	*How many departments are part of Overseas France?*
07	Ordinal	<x> <rdf:type> <X> <x> <DP> <Value> **ORDER BY DESC**(Value) **OFFSET 1 LIMIT 1**	Which <X> has the 2nd most <Value>?	*Which city has the 2nd highest population?*
08	Condition	<x,y> <subClassOf> <Z> <x> <DP> <value1> <y> <DP> <value2> **FILTER** (value1 = value2)	Which <Z> share the value of <DP>?	*Which cities does the Seine River pass through?*

for learners to find the answer. Other studies [4, 7] explored various ontological measures to extract semantic features, such as using entity popularity as a determining factor of difficulty, as questions containing popular entities were believed to be easier to answer. To measure this, the authors counted the number of object properties linked to the entity from other individuals within the ontology. They also proposed measuring question specificity by examining the depth of a certain concept in the concept and role hierarchy of the domain ontology, as deeper concepts in the hierarchy result in more difficult questions. Other work [10, 18] proposed similar features, but used a knowledge base to extract these features.

The literature suggests that research on automated evaluation methods is constantly evolving; however, certain approaches are limited by factors such as question types [5] or the specific attributes of the input ontology [4]. Challenges still exist in developing comprehensive evaluation frameworks that can accommodate various question types and ontology structures.

3. The Proposed Framework

3.1. Characteristics of Questions Generated from Ontologies

The generation of questions from ontologies typically requires a set of templates (textual or graph-based) that are instantiated with ontological elements based on rules. The basic building block behind the instantiation is the existence of Resource Description Framework (RDF) patterns that facilitate the generation of certain question types. Different templates require the ontology to contain specific RDF patterns as pre-requisites to generate a required question.

Several of the most common question formats are illustrated in Table 1, together with the corresponding RDF pattern requirements used in other studies [13, 19, 20], which may vary in terms of the number of triple patterns they require and the types of properties they include.

RDF patterns may involve the utilisation of concepts [5, 20], various types of properties [13], individuals [20], quantifiers [19] or constraints[1] [16, 21]. Additionally, they may include a single triple pattern or multiple triple patterns. For example, the first template in Table 1 is used to generate definition questions where the learner is asked to define a certain domain-specific term. As a pre-requisite, the ontology must contain its corresponding RDF pattern which necessitates the existence of an entity of type `Class` that is annotated with a comment using `<rdfs:comment>`. Another template requires the incorporation of entities that are connected through object- or datatype properties to generate questions that ask about the relationship between two entities. Such questions have a one-to-one mapping between the number of relevant triples and generated questions which indicate that these templates rely on explicitly stated facts. For example, if the ontology includes the triple `<Paris, capitalOf, France>`, the system may generate a question based on the *Property Assertions* template (Table 1) *"Is Paris the capital of France?"*. This approach depends on utilising the information available in the ontology directly, without any additional inference.

Other templates may include functional properties that impose constraints on the question [16] to derive new knowledge and generate questions that inquire about facts that are not explicitly modelled by the ontology. Functional properties allow the system to infer new facts that are not explicitly mentioned in the ontology. The constraint-based question in template 07 requires the existence of multiple entities that are connected through a specific datatype property and applies an ordinal constraint (represented by ordinal or relational operators) to generate the questions. For example, if the input ontology includes information about the cities in France and their respective populations, the model can generate a question such as *"Which city in France has the second highest population?"* (example question for template 07 in Table 1). This enhances the capability of question generation by producing more complex questions.

3.2. Theoretical Foundation

There are different perspectives in the pedagogical literature on what constitutes a complex question. Complexity is often seen as a measure of the cognitive demand placed on the learner by the question. Ahmed and Pollitt [22] explored the cognitive demands of questions, and suggested that question complexity is a crucial aspect that increases cognitive demand. They defined complexity as the number of operations or ideas that need to be considered in order to arrive at a correct answer. Low-complexity questions involve straightforward ideas (i.e. recognition of knowledge) and operations that do not require linking them together, whereas higher-complexity questions require the learner to identify and combine various operations and concepts and connect them, hence promoting recall and evaluation of knowledge. Studies examining the gradual progression towards competence found that novice and expert learners possess different characteristics regarding their level of knowledge and ability to reason about that knowledge [23, 24]. These studies suggest that assessments need to consider this distinction

[1] A constraint is defined as a triple in which the subject and object are connected through a *functional property*.

in order to distinguish between learners of different mastery levels by constructing questions that require varying amounts of knowledge and different reasoning abilities. This relates to the previous definition of complexity, which defines complex questions as those that encompass a greater number of facts and necessitate higher cognitive skills, such as deduction and reasoning. To illustrate this, in order to correctly answer the example question for template 07 in Table 1, a learner needs to understand several semantic relations (i.e. inference steps): 1) the answer must be a city in France; 2) the answer should have a numerical value indicating its population; and 3) the learner must select the answer that ranks second in terms of population among all cities. This contrasts with the example question for template 03 that requires the learner to recall a single piece of information (i.e. *"The capital city of France is Paris"*). For a more detailed discussion of this theoretical backing, see [25].

3.3. Question Complexity Evaluation Framework

The main objective of our framework is to assess the complexity of questions by utilising ontological metrics that are shared by questions generated from ontologies. This allows us to distinguish between students with varying levels of mastery. Complexity, in this context, refers to the level of knowledge required, and the cognitive demands placed on students during the question-answering process. Instead of heavily relying on educational experts to determine the complexity level of the generated questions, we suggest utilising the ontological features that make up the questions to automatically calculate the internal complexity of the generated questions. By *internal*, we mean the intrinsic factors of the question that increase or decrease its complexity level. Thus, we eliminate external sources such as the proficiency level of learners or previous background knowledge. We evaluate the generated questions based on two metrics (based on those described in [25]): 1) the volume of knowledge (i.e. number of facts) they test; and 2) the level of reasoning they require. The first metric quantifies the number of relevant triples used to generate the question; note that this metric does not count the number of entities in the question itself (as assessment questions are short linguistic constructs which typically contain a limited number of facts) but rather it establishes a relationship between the number of relevant triples retrieved from executing the query against the ontology (during instantiation) and the number of generated questions. The second metric evaluates the specificity and restrictiveness of the question by quantifying the number of applied constraints. This metric indicates the desired level of precision in the answer and highlights the question's role in filtering and refining the search space. We use this metric to determine the level of reasoning involved in generating the answer. The proposed metrics are calculated according to the following definition:

Definition: Let Q be a question, GP be its corresponding Graph Pattern, CGP be the complexity of the Graph Pattern, TPi be a triple pattern where each element (subject, predicate, and object) can be a variable and C be a constraint. The complexity of a question Q is equal to the complexity of its corresponding graph pattern CGP, and the complexity of the graph pattern CGP is calculated as the total number of triple patterns and constraints it contains.

Thus, template 03 has a complexity score of 1, as it only has one relevant triple and no constraints, whereas template 07 has a higher complexity, due to the three constraints ('`ORDER BY DESC`(Value)', '`OFFSET 1`', and '`LIMIT 1`') and the additional triple involved. These metrics enhance evaluation methodologies in the field of ontology-based AQG, thus enabling a more

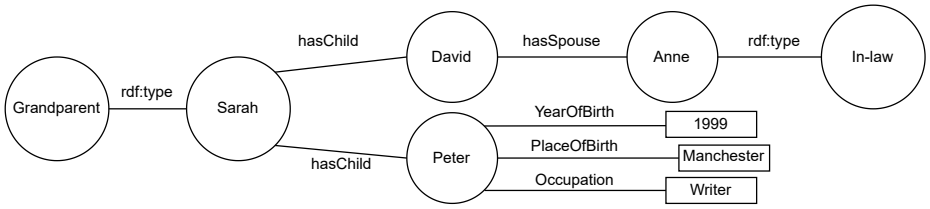


Figure 1: A fragment of the family ontology illustrating some family members and their characteristics.

comprehensive and advanced assessment of the generated questions' complexity level.

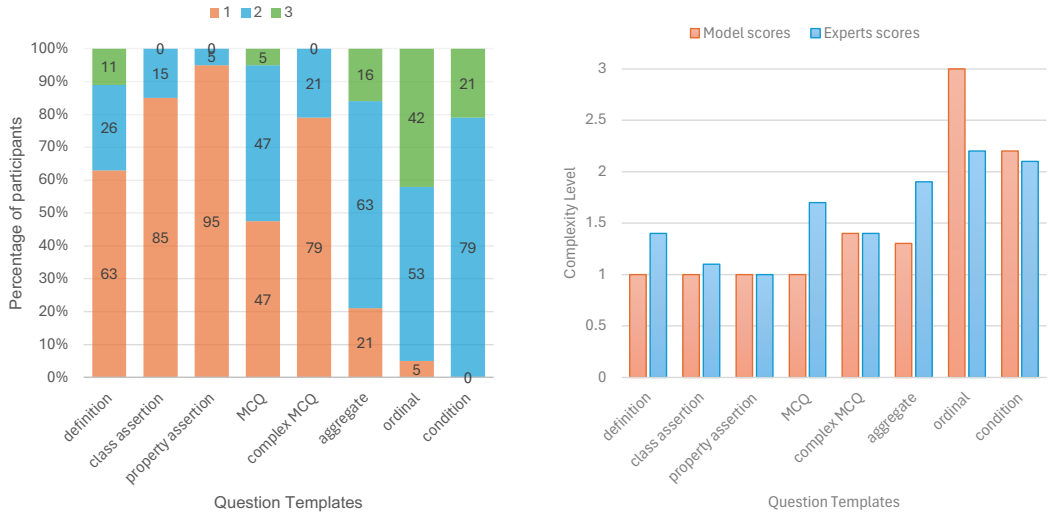
4. Evaluation

4.1. Methodological Approach

An expert-centred evaluation was conducted where educational experts with knowledge and experience in the construction of assessment questions in an educational setting were recruited. They were presented with a set of generated questions with varying characteristics and asked to rate each question based on its complexity level according to their perception. The questions can be rated on a scale ranging from 1 to 3, where a higher score denotes higher complexity. The evaluation was conducted through an online questionnaire designed for this purpose. As the data collection for this evaluation is currently ongoing, the results presented here are preliminary, and provide an initial assessment of the proposed evaluation framework.

4.2. Experimental Preparation

Our study utilised a domain ontology purposely built for this evaluation task. Instead of using an ontology focused on a specific domain, we opted to model a universally recognised concept: the family tree relationship. This decision was made to provide a foundation of abstract knowledge that anyone can understand, regardless of their expertise or familiarity with specific subjects. By relying on the inherent understanding that people from different cultures and backgrounds share, we aim to create a neutral basis for assessing question complexity. This approach allows us to evaluate questions consistently and ensures that their quality is determined solely by their structure and cognitive attributes, without being influenced by domain-specific complexities. A small ontology was therefore developed (comprising 130 axioms) representing information about an imaginary family including the relationship between each individual and some characteristics including individuals' year of birth, place of birth, I.Q. and occupation. A fragment of the input ontology is illustrated in Figure 1. The question generation process was guided by the question specifications shown in Table 1, which allowed us to generate questions of varying characteristics, based on those proposed in previous studies. Multiple variants were generated from each template, and in order to keep the rating process manageable, we randomly selected and presented 3 variants per template to our experts. Consequently, each expert evaluated a total of 23 questions.



(a) Expert-based assessment of complexity for different generated questions.

(b) A comparison of complexity scores generated by our model with expert perceptions.

Figure 2: Comparison of expert-based assessments of generated questions, and the complexity model.

5. Preliminary Results

The preliminary analysis revealed notable trends regarding the perceived complexity of the questions. Figure 2a shows the percentage distribution of the ratings on a 3-point scale of complexity, whereas Figure 2b presents a comparison between the complexity scores generated by our model and those given by the experts. To obtain the automatic scores, we calculated the complexity of each variant that was generated from a given template, and from these, determined the mean rating given for the template. Finally, the complexity levels were normalised to map values to a range between 1 and 3 for better comparability with experts' ratings, based on the minimum and maximum values of complexity given the dataset of questions considered.

Both figures suggest some correlation between the complexity scores provided by experts and those generated by the model, indicating consistency in evaluating question complexity through both approaches. The majority of the respondents' ratings closely matched the assessments made by our framework, with only a few minor discrepancies. More specifically, questions that our framework categorised as simple (templates 01 to 03, excluding template 04) were also perceived as simple by most reviewers. Regarding *Definition*, *Property Assertion*, and *Class Assertion* questions, our model assigned a complexity rating of 1, implying a level of simplicity. The experts' ratings strongly aligned with our system's ratings, with approximately 98% of these questions also receiving a complexity rating of 1 from the experts. Questions based on *Property Assertions* were perceived to be the simplest by the experts across those questions analysed as having a very low complexity, with a mean very close to 1. This closely aligned with the score given by our model, given that the answer to these questions is encoded within a single triple pattern (i.e. <x> <p> <y>) without requiring any additional cognitive tasks.

The second template, which our model considered as producing simple questions, generated questions based on class assertion axioms using the template `<x> <rdf:type> <X>`, with 85% of experts considering the questions generated from this template to be simple, with a mean rating close to 1. Additionally, 15% of experts gave these questions a maximum score of 2.

The third template, which generates definition questions, differed slightly from the previous templates in terms of simplicity according to our model. While still perceived as low complexity, this template exhibits more variability in responses. The mean value is higher than 1, indicating that some experts found it moderately complex. The distribution shows that approximately 60% of experts rated it as low complexity, while medium and high ratings accounted for the remaining 40%. This observation can be attributed to several factors. Firstly, the answer space for definition questions is typically broader, allowing for more potential variations in responses. This requires greater precision from the expert and may involve other skills, such as linguistic proficiency. When examining the specific textual templates used for this template, we utilised two different variations. These variations are "`Define <X>`" and "`What is the term used to describe this <string>?`" The question generated from the first textual template was perceived as more complex than the second, possibly due to the cognitive tasks involved in defining a term, such as categorisation and abstraction, which are mentally demanding compared to simply recognising or identifying a term. This supports the notion that the expanded answer space could influence the perception of complexity. Thus, these factors contribute to the understanding that definition questions are not as straightforward as other types of questions.

While our system categorised *MCQs* as simple, they were perceived as moderately complex, with a mean below 2. The distribution of complexity ratings is balanced between low and medium complexity, with very few high ratings. Upon closer examination,

we discovered that our model only considered the triple patterns present in the stem "`Which of these is <X>:`" when calculating the complexity score for MCQs, and excluded those that appeared in the options. However, MCQs were generated and presented to the experts with the options included in the stem (e.g., "`Which of these is X: 1)y 2)x 3)z`"). This omission caused some triples to be excluded from the complexity calculation, resulting in a lower score for the number of relevant triples metric and an overall decrease in the complexity score. This finding has prompted us to re-evaluate our model for the MCQ template in order to address the issue of missing triples.

The template used to create *Complex MCQs* aligns with the score given by our model. Our model rated this template with a complexity score closer to 1.5, indicating moderate complexity. This is due to the larger number of triples present in the questions generated from this template. Experts also gave the question a score of 1.5, indicating moderate overall agreement. The distribution reveals that while the majority of ratings are for low complexity, with some for medium complexity, no high ratings were given to the question. This suggests that the question leans towards simplicity.

Based on the expert evaluations, templates 06 (*Aggregate*), 07 (*Ordinal*), and 08 (*Conditional*) were considered the most complex. Of these, our model's ratings align with this assessment for templates 06 and 07. However, when it comes to questions generated based on *Aggregate* functional properties, our model's ratings differed the most from those given by experts.

Only 20% of experts considered *aggregate*-based questions to be of low complexity. The majority of experts, 63%, gave these questions a score of 2, whereas 16% gave them a score

of 3. Overall, these questions were perceived to be on the complex side, with a mean closer to 2. However, our model assigned these questions a score closer to 1, despite the fact that the question template included multiple triple patterns and utilised constraints such as COUNT and HAVING, which introduce an additional reasoning step in the question-answering process. This discrepancy becomes clearer when we compare *aggregate* questions to *complex MCQs*. Our model assigned a higher complexity score to the latter, despite the fact that they do not require any reasoning. This highlights the importance of revising the definition of complexity and possibly giving more weight to the second metric: *the number of constraints included*. This suggests that the use of weights for different constraints may result is a more nuanced complexity estimate, which could be explored in future work. By making this adjustment, our model would be able to differentiate between questions with multiple triple patterns but no reasoning, and questions that utilise both features.

The templates that received the fewest low ratings were the *Ordinal* and *Condition* templates, when compared to other questions. *Ordinal* questions were found to be the most complex, with a mean score of 2.2. However, it is worth noting that our framework categorises these questions as the most complex, while the automatic score rates them even higher at 3, which is considerably more complex than that judged by the experts. Although the complexity ratings were spread fairly evenly between medium and high, there were very few low ratings for this template. Similarly, none of the experts rated questions generated from the *Condition* templates as having low complexity, as there were no low ratings in the distribution. The majority of ratings for this template were medium complexity, with some high ratings, which aligns with the score given by our model suggesting an overall score closer to 2. This suggests that experts widely agreed that *Ordinal* and *Conditional* questions should not be regarded as simple questions (with respect to complexity).

Overall, these initial findings support the hypothesis that an automatic evaluation framework can be effectively used as a proxy for expert user evaluations when assessing the complexity level of questions generated from ontologies. However, they also identify areas where the model could be refined, especially in accurately evaluating the complexity of MCQs and considering the added complexity introduced by constraints. We will continue to analyse and refine our evaluation approach in future work to better align our system's ratings with expert assessments.

6. Conclusion

This paper introduced an automatic framework that leveraged ontological metrics to assess the complexity of questions generated from domain ontologies. The effectiveness of the proposed framework was evaluated through an expert-based evaluation which revealed consensus between the complexity scores generated by the framework and the opinions of educational experts. Nonetheless, minor discrepancies arose, particularly in scenarios involving questions with more triples and those requiring higher levels of reasoning. These findings highlighted the need for adjustments to address these discrepancies. This study contributes to ongoing research on automatic evaluation methods for generated questions, which complements traditional approaches and improves scalability.

References

[1] R. Studer, V. R. Benjamins, D. Fensel, Knowledge engineering: Principles and methods, Data & knowledge engineering 25 (1998) 161–197.

[2] S. AlKhuzaey, F. Grasso, T. R. Payne, V. Tamma, Text-based question difficulty prediction: A systematic review of automatic approaches, International Journal of Artificial Intelligence in Education (2023) 1–53.

[3] L. Benedetto, P. Cremonesi, A. Caines, P. Buttery, A. Cappelli, A. Giussani, R. Turrin, A survey on recent approaches to question difficulty estimation from text, ACM Computing Surveys 55 (2023) 1–37.

[4] E. V. Vinu, P. S. Kumar, A novel approach to generate MCQs from domain ontology: Considering DL semantics and open-world assumption, Journal of Web Semantics 34 (2015) 40–54.

[5] T. Alsubait, B. Parsia, U. Sattler, Ontology-based multiple choice question generation, KI-Künstliche Intelligenz 30 (2016) 183–188.

[6] J. Leo, G. Kurdi, N. Matentzoglu, B. Parsia, U. Sattler, S. Forge, G. Donato, W. Dowling, Ontology-based generation of medical, multi-term mcqs, International Journal of Artificial Intelligence in Education 29 (2019) 145–188.

[7] V. E. Venugopal, P. S. Kumar, Difficulty-level modeling of ontology-based factual questions, Semantic Web 11 (2020) 1023–1036.

[8] L. Zhang, K. VanLehn, How do machine-generated questions compare to human-generated questions?, Research and practice in technology enhanced learning 11 (2016) 1–28.

[9] C. Lin, D. Liu, W. Pang, E. Apeh, Automatically predicting quiz difficulty level using similarity measures, in: Proceedings of the 8th international conference on knowledge capture, 2015, pp. 1–8.

[10] A. Faizan, S. Lohmann, V. Modi, Multiple choice question generation for slides, in: Computer Science Conference for University of Bonn Students, 2017, pp. 1–6.

[11] A. Faizan, S. Lohmann, Automatic generation of multiple choice questions from slide content using linked data, in: Proceedings of the 8th international conference on web intelligence, mining and semantics, 2018, pp. 1–8.

[12] C. Jouault, K. Seta, Y. Hayashi, Quality of lod based semantically generated questions, in: Artificial Intelligence in Education: 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings 17, Springer, 2015, pp. 662–665.

[13] K. Stasaski, M. A. Hearst, Multiple choice question generation utilizing an ontology, in: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, 2017, pp. 303–312.

[14] G. Kurdi, B. Parsia, U. Sattler, An experimental evaluation of automatically generated multiple choice questions from ontologies, in: OWL: Experiences and Directions–Reasoner Evaluation: 13th International Workshop, OWLED 2016, and 5th International Workshop, ORE 2016, Bologna, Italy, November 20, 2016, Revised Selected Papers 13, Springer, 2017, pp. 24–39.

[15] F. B. Baker, S.-H. Kim, et al., The basics of item response theory using R, volume 969, Springer, 2017.

[16] S. Alkhuzaey, F. Grasso, T. R. Payne, V. Tamma, Generating complex questions from

ontologies with query graphs, 2024. Proceedings of the 28th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES).
[17] C. Jouault, K. Seta, Y. Hayashi, Content-dependent question generation using lod for history learning in open learning space, New Generation Computing 34 (2016) 367–394.
[18] D. Seyler, M. Yahya, K. Berberich, Knowledge questions from knowledge graphs, in: Proceedings of the ACM SIGIR international conference on theory of information retrieval, 2017, pp. 11–18.
[19] T. Raboanary, S. Wang, C. M. Keet, Generating answerable questions from ontologies for educational exercises, in: Research Conference on Metadata and Semantics Research, Springer, 2021, pp. 28–40.
[20] B. Diatta, A. Basse, S. Ouya, Bilingual ontology-based automatic question generation, in: 2019 IEEE Global Engineering Education Conference (EDUCON), IEEE, 2019, pp. 679–684.
[21] J. Bao, N. Duan, Z. Yan, M. Zhou, T. Zhao, Constraint-based question answering with knowledge graph, in: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers, 2016, pp. 2503–2514.
[22] A. Ahmed, A. Pollitt, Curriculum demands and question difficulty, in: IAEA conference, Bled, Slovenia, 1999.
[23] M. T. Chi, P. J. Feltovich, R. Glaser, Categorization and representation of physics problems by experts and novices, Cognitive science 5 (1981) 121–152.
[24] M. T. Chi, R. D. Koeske, Network representation of a child's dinosaur knowledge., Developmental psychology 19 (1983) 29.
[25] S. Alkhuzaey, F. Grasso, T. R. Payne, V. Tamma, A framework for assessing the complexity of auto generated questions from ontologies, in: Proceedings of the European Conference on e-Learning, volume 22, 2023, pp. 17–24.