Generating Answerable Questions from Ontologies for Educational Exercises

Toky Raboanary[0000−0001−6133−4643], Steve Wang, and C. Maria Keet[0000−0002−8281−0853]

Department of Computer Science, University of Cape Town, South Africa
traboanary@cs.uct.ac.za, WNGSHU003@myuct.ac.za, mkeet@cs.uct.ac.za

Abstract. Proposals for automating the creation of teaching materials across the sciences and humanities include question generation from ontologies. Those efforts have focused on multiple-choice questions, whereas learners also need to be exposed to other types of questions, such as yes/no and short answer questions. Initial results showed it is possible to create ontology-based questions. It is unknown how that can be done automatically and whether it would work beyond that use case in biology. We investigated this for ten types of educationally useful questions with additional sentence formulation variants. Each type of questions has a set of template specifications, axiom prerequisites on the ontology, and an algorithm to generate the questions from the ontology. Three approaches were designed: template variables using foundational ontology categories, using main classes from the domain ontology, and sentences mostly driven by natural language generation techniques. The user evaluation showed that the second approach resulted in slightly better quality questions than the first, and the linguistic-driven templates far outperformed both on syntactic and semantic adequacy of the questions.

Keywords: Ontology-based Question Generation · Ontologies for education · Natural Language Generation.

1 Introduction

Ontologies and knowledge graphs are used in an increasing variety of ontology-driven information systems. Our focus is generating questions from ontologies for educational purposes. If there is an annotated textbook in cultural heritage, one can link it to an ontology and develop an educational game by generating educational questions to foster active learning in the same spirit as alluded to in [6]. Question generation from an ontology or linked data has been investigated mainly for multiple-choice questions (MCQs) using tailor-made algorithms or SPARQL queries [2, 20, 23], knowledge graph construction for it [21], and architectures more broadly [22]. There are multiple types of questions beyond MCQ, such as similarity, yes/no, and short answers that may be automatically marked as well [6, 21]. Here, we focus on the two latter types of questions. For instance, from the axiom $\mathsf{Collection} \sqsubseteq \forall \mathsf{hasMember}.(\mathsf{Collection} \sqcup \mathsf{CulturalHeritageObject})$ in

Cultural-On [11], one could generate a question "Does a collection have a member that is only a cultural heritage object?". This opens up many possibilities for question construction for multiple axiom types, as well as combinations thereof; e.g., given $\mathsf{CulturalInstituteOrSite} \sqsubseteq \forall \mathsf{isSubjectOf}.\mathsf{CreativeWork}$ and $\mathsf{CulturalInstituteOrSite} \sqsubseteq \mathsf{CulturalEntity}$, to generate "Which cultural entity is a subject of only a creative work?". It is unclear what the prerequisites of the ontology are, i.e., which axiom(s) type(s) is (are) needed for which type of educational questions, and which type of questions one possibly could generate from an ontology. Questions can be generated from instance or type-level information (ABox or TBox), where we zoom in on the TBox since it is relevant for learning generic knowledge. In this paper, we aim to answer the following questions:

1. Which of the types of questions that are educationally relevant can be generated from the TBox of an ontology? Or, from the ontology viewpoint: What are the axiom prerequisites, i.e. types of axioms that must be in the ontology, to be able to generate a particular type of educational question?
2. Can the outcome be generalised to any combination of ontology (+ textbook) with question templates whilst maintaining good quality questions?

We aim to answer these questions in this paper. Taking the principal types of questions as identified by education research, we systematically assess what the axiom prerequisites are and devise templates for the questions with linguistic variants. A template is a linguistic structure containing gaps that are intended to be filled in to create a sentence. We examined 10 educational types of questions and their axiom prerequisites, represented in the description logic $\mathcal{ALC}$. Three different approaches were developed and implemented to automatically generate the questions from the ontology: 'basic' templates with DOLCE [15] categories for key variables, templates that use a top-level vocabulary of the domain ontology to tailor the basic templates, and natural language generation (NLG)-based tailoring of the basic templates, where the first two approaches informed the third one. The generated questions were evaluated by humans on perceived syntactic and semantic correctness. The first two approaches resulted in poor performance (26% and 34% of good quality), whereas the domain-independent but NLG-enhanced templates reach over 80% very good syntactic and 73.7% as good or very good semantic quality. The algorithms, source code, templates, generated questions, ontologies and data used in the experiment are available at `https://github.com/mkeet/AQuestGO`.

The remainder of the paper is structured as follows. We present the related work in Section 2, the question generation in Section 3, and the evaluation with discussion in Section 4. We conclude in Section 5.

2 Related work

Questions can be generated from ontologies [2, 6, 7, 19, 22, 23], using either generic systems [2, 7, 19, 23] or tailor-made for a specific domain, such as biology [6, 24] and mathematics [14]. They may have a new purposely-built [6] or existing [7, 23] ontology as input. Most research focuses on MCQ generation [2, 7, 19, 23], which mainly deal with distractor generation and difficulty control.

Concerning the verbalisation, i.e., generating the natural language sentences, only [5, 24] evaluated the linguistic quality of the generated questions. Bühmann et al. [5] considered their syntax (fluency) and their semantics (adequacy), but the sentences are over the ABox rather than the TBox. Zhang and VanLehn [24] evaluated the fluency and ambiguity of their questions, but their approach is designed for one knowledge base. Vinu et al. [23] consider the surface structure of generated questions with regex, yet they did not evaluate their verbalisation approach. Also, the generalisability of approaches is found wanting: most of them used only one ontology in their experiment, except those which used three [1, 7] and four [23] ontologies.

Chaudhri et al.'s idea for non-MCQ educational question generation with their "intelligent textbook" [6] is appealing for fostering active learning. However, they did not make their question templates or the construction process available, nor is it clear how this could be reused for other ontologies beyond their "Inquire Biology" use case for one hand-crafted ontology and one particular textbook.

Question generation is also used for other tasks; notably, ontology validation [1]. Abacha et al. [1] evaluated their questions, but covered only a subset of possible sentence constructions, such as omitting quantifiers explicitly. Further afield, there are statement generation verbalisation systems [4], and frameworks [17] for verbalising RDF, OWL and SPARQL, whose experiences may be of use, but they do not generate (educational) questions.

3 Question generation

The design choices are described before we proceed to the question specifications and algorithms.

3.1 Design choices

There are core choices for the template design within the context of ontology-based question generation in anticipation of their quality. For the templates themselves, there are four core options:

Type A: Fixed template structure where one fills in the slots with the relevant variable (class, object property (OP), quantifier) fetched from the ontology, at that level of specification; e.g., *Is a* [owl:thing] [owl:objectproperty] [quantifier] [owl:thing]? as template which could have an instantiation resulting in, e.g., "Is a cultural heritage object a member of some collection?".

Type B: As Type A, but specify the category at least, especially for the OWL class; e.g., that it has to be a dolce:process, or a bfo:continuant (cf. owl:thing), so that for the template instantiation, it will pick that or any of its subclasses so as to broadly constrain the filler type. This is likely to increase the quality of the syntax and semantics of the generated questions. A foundational ontology is well-suited for this.

Type C: As Type B, but tailor the template with the domain ontology vocabulary to some degree; e.g., select a high-level class from the domain ontology, e.g., CulturalEntity from Cultural-On, so that the considered slot of the template will only be instantiated with a subclass of culturalon:CulturalEntity. One may expect better semantics of the questions, but it comes at the cost of reduced generalisability across domain ontologies.

Type D: Contextualise the templates based on the ontology vocabulary using NLG techniques, but do not perform tailoring of slots with any ontology vocabulary. This assumes that the question quality is more dependent on the linguistic realisation module of the NLG process than on the representation of the domain knowledge.

3.2 Types of questions and their prerequisites

The types of questions considered in this paper are adjusted from [6] and extended with questions from the Webclopedia QA typology [10] that is based on actual educational questions. They are also included in [9] and are shown to be suitable for education [18]. We chose this typology because its question templates are abstract (not domain-specific), which is appropriate for the generalisability purpose, and it is based on 17,384 questions and their answers.

Templates of different question types are specified, and each slot in the template is replaced by the appropriate class or object property (OP) or quantifier in an ontology. We selected DOLCE [15] for the Type B templates, but one could take another foundational ontology. For the Type C examples below, terms in Cultural-On are used. Each question template is mapped to Description Logic (DL) queries to check that the generated question is answerable by the ontology. For Type D, we devised several templates (e.g., templates in active/passive voice and hasX OP naming format) for each type of questions.

The aggregate number of variants of templates designed for the three approaches are presented in Table 1. The different numbers of variants are due to peculiarities of the approaches, such as more tailoring with domain ontology vocabulary (hence |Type A/B| ≤ |Type C|), and accommodating active/passive voice or not. Due to space limitations, we present all types of questions with their prerequisites only briefly and more details can be found online.

Yes/No and True/False Questions These questions expect yes/no or true/false as an answer. Since the ontology operates under Open World Assumption, the answer to a question is no only if the ontology explicitly states so. For instance, using Thing or any of its subclasses, a template "*Does a* X OP *a* Y*?*" (for numbers *i,iv* in Table 1) can be generated if $\mathsf{X} \sqsubseteq \exists \mathsf{OP}.\mathsf{Y}$ or $\mathsf{X} \sqsubseteq \forall \mathsf{OP}.\mathsf{Y}$ (Answer: Yes) or if $\mathsf{X} \sqsubseteq \neg \forall \mathsf{OP}.\mathsf{Y}$ (Answer: No). Template examples of this type are:
Type A template: *Does a* [Thing] [OP] *a* [Thing]*?*
Type B template: *Does a* [Endurant] [OP] *a* [Thing]*?*
Type C template: *Does a* [CulturalEntity] [OP] *a* [Thing]*?*
Type D templates: *Does a* [T_Noun] [OP_Verb] *a* [T_Noun]*?*
Does a [T_Noun] [OP_Verb_Prep] *a* [T_Noun]*?*
where for Type D, T_Noun states that the class name Thing is a noun, OP_Verb means that the OP name is a verb and OP_Verb_Prep indicates it also has a preposition. Then, "*A* X OP *some* Y. *True or false?*" *(ii,v)* and "*A* X OP *only*

Table 1: Numbers of variants of templates by type of template.

Group of TQ	**No.**	**Type of Questions (TQ)**	**A/B**	**C**	**D**
Yes/No	*i*	Two classes and one property	4	6	6
	ii	Two classes, one property, and a quantifier	4	4	10
	iii	One Endurant and one Perdurant	4	4	1
True/False	*iv*	Two classes and one property	4	6	10
	v	Two classes, one property, and a quantifier	4	6	20
Equivalence	*vi*	Equivalence	2	5	3
Subclass	*vii*	Two classes and one property	1	4	5
	viii	Additional quantifier	1	1	10
	ix	One class and one property	4	4	4
Narrative	*x*	Narrative	2	2	6
		Total	30	42	75

a Y. *True or false?*" *(ii,v)* can be generated if $\mathsf{X} \sqsubseteq \exists \mathsf{OP.Y}$ (Answer: Yes) or if $\mathsf{X} \sqsubseteq \neg\exists \mathsf{OP.Y}$ (Answer: No), and if $\mathsf{X} \sqsubseteq \forall \mathsf{OP.Y}$ (Answer: Yes) or if $\mathsf{X} \sqsubseteq \neg\forall \mathsf{OP.Y}$ (Answer: No), respectively. Finally, "*Does a* X Y*?*" *(iii)* can be generated if $\mathsf{X} \sqsubseteq \exists \mathsf{participates\text{-}in.Y}$ (Answer: Yes), or if $\mathsf{X} \sqsubseteq \neg\exists \mathsf{participates\text{-}in.Y}$ (Answer: No).

Equivalence Questions This is possible to generate provided the two classes are asserted or inferred to be equivalent. The template "*Are there any differences between a* X *and a* Y*?*" (*vi* in Table 1) can be generated and results in "Yes" if $\mathsf{X} \equiv \neg\mathsf{Y}$, and "No" if $\mathsf{X} \equiv \mathsf{Y}$ is asserted or inferred in the ontology.

Subclass Identification Questions These questions can be casted as "Which" questions. The template "*Which* X OP Y*?*" *(vii)* can be generated if there is a class Z that satisfies the axiom pattern $\mathsf{Z} \sqsubseteq \mathsf{X} \sqcap \exists \mathsf{OP.Y}$ or $\mathsf{Z} \sqsubseteq \mathsf{X} \sqcap \forall \mathsf{OP.Y}$. Then, the template "*Which* X OP *some* Y*?*" *(viii)* can be generated if there is a class Z that satisfies the axiom pattern $\mathsf{Z} \sqsubseteq \mathsf{X} \sqcap \exists \mathsf{OP.Y}$. The template "*Which* X OP *only a* Y*?*" *(viii)* can be generated if there is a class Z that satisfies the axiom pattern $\mathsf{Z} \sqsubseteq \mathsf{X} \sqcap \forall \mathsf{OP.Y}$. Finally, "*What does a* X OP*?*" *(ix)* can be generated if there is a class Y such that $\mathsf{X} \sqsubseteq \exists \mathsf{OP.Y}$ or $\mathsf{X} \sqsubseteq \forall \mathsf{OP.Y}$.

Narrative Questions A class X in an ontology can be "defined" if it satisfies one of the following criteria: 1) it is annotated with a definition; 2) it has at least one equivalent class; 3) it has at least one superclass, at least one subclass or a combination of both; for instance, "*Define* X." (number *x* in Table 1).

The 10 types of educational questions with their specific axiom prerequisites presented as a summary here answer our first research question. The full specifications can be found in the supplementary material online.

3.3 Dynamic question generation: the algorithms

This section presents an overview of the three approaches we have designed for the dynamic question generation: template variables using foundational ontology categories (Appr 1), using main classes from the domain ontology (Appr 2), and sentences mostly driven by natural language generation techniques (Appr 3). Appr 1 and Appr 2 adopt 'Algorithm 1', with the difference that the former takes Type A and Type B templates as input and the latter takes Type C templates as input. Appr 3 uses ‘Algorithm 2’ and takes Type D templates as input. All details about the algorithms can be found in the supplementary material.

Algorithm 1: ontology element-based templates Algorithm 1 is composed of some variant sub-algorithms depending on the type of questions, but several steps are the same. There are 3 different types of tokens that are going to replace the slots in templates: quantifier tokens (denoted with [quantifier]), OWLObjectProperty tokens, and OWLClass tokens. A [quantifier] in the template is replaced with either ‘some’ ($\exists$) or ‘only’ ($\forall$). When the token appears as an [ObjectProperty] then it can be replaced with any of its object subproperties in the ontology that satisfies the axiom prerequisites of the question type. If [X], indicating an OWLClass, appears in the template, then it can be replaced with any subclass of X.

Overall, the algorithm picks a template and tries to fill it with contents from the ontology, taking into account the vocabulary, axiom prerequisites, hyphen checking (e.g., ‘Bumble-Bee’ is converted to ‘bumble bee’) and article checking (e.g., ‘a elephant’ is converted to ‘an elephant’). For example, with the template “*Does a* [Thing] [ObjectProperty] *a* [Thing]*?*”, the algorithm can generate a question like “Does a catalogue describe a collection?” from the axiom $\mathsf{Catalogue} \sqsubseteq \exists \mathsf{describes.Collection}$.

Algorithm 2: natural language-driven templates Algorithm 2 not only fills in the question templates, but also fetches all axioms satisfying the axiom prerequisites from a selected type of questions. Then, it processes the contents of the ontology by fetching the vocabulary elements of a selected axiom, picks an appropriate variant of a template that the vocabulary can be used in, and makes some linguistic adaptation before generating the whole question.

The improvements incorporated were partially informed by the analysis of the ‘bad’ sentences generated by Algorithm 1. There are three major changes:

- *using class expressions to generate questions, rather than only the declared domain and range of OPs*, so using only asserted and inferred knowledge.
- *improving common grammar issues, availing of SimpleNLG [8] and WordNet [16], for subject and verb agreement, gerund form generation, and article checking.* Also, a basic part of speech (POS) tagging for the classes and OPs was added to get the appropriate form, by using WordNet [16].
- *choosing the appropriate template for a given axiom by considering the POS of classes and OPs, and classifying the OP.* We designed an algorithm based on an FSM that classifies the name given to an OP to find the appropriate template for an axiom and provides the appropriate equivalent text. It considers 6 linguistic variants. An OP name may: 1) have a verb, 2) start with a verb followed by a preposition, 3) start with ‘has’ and followed by nouns, 4) be composed of ‘is’, nouns and a preposition, 5) start with ‘is’, followed by a verb in a past participle form and ends with a preposition, or 6) start with ‘is’, followed by a verb in a past participle form and ends with ‘by’ (i.e., passive voice variants for 4-6). The FSM strategy is a sequence detector to determine the category of an OP and chunks it. For instance, the OP

is-eatenBy, which is an instance of OP_Is_Past_Part_By (the 6^{th} variant), is transformed into a list of words (is, eaten, by). Then, it detects each component, and from that, the POS of each token is obtained, and, finally, it generates the appropriate group of words: "is eaten by", which will be used in the question.

So, for the axiom $\mathsf{Leaf} \sqsubseteq \exists \mathsf{eaten\text{-}by.Giraffe}$, the appropriate template is "*Is a* [T_Noun][OP_Is_Past_Part_By] *a* [T_Noun]*?*" and a correct generated question would be "Is a leaf eaten by a giraffe?" rather than "Does a leaf eaten by a giraffe?". Finally, the mapping between the vocabulary elements of the axiom and the tokens of the selected template is done sequentially.

4 Evaluation

The evaluation aims to gain insight into the quality of the algorithms with respect to 1) the syntax, 2) the semantics of the sentences, and 3) the generalisability of the approach to multiple ontologies. To this end, we have conducted three evaluation sessions. The first two evaluation sessions with Appr 1 and Appr 2, using Algorithm 1 were of a preliminary nature, in that we focused only on the first two aims of the evaluation and used only one ontology. The third evaluation also considered the potential for generalisability using Appr 3 with Algorithm 2. Ethics approval was obtained before the evaluation sessions.

4.1 Materials and Methods

Materials Three ontologies were used in our evaluation: an extended version of African Wildlife Ontology (AWO) [13], where we added 19 classes (a.o., BumbleBee, Land, Fly) and 4 OPs (a.o., participate-in, live-on) so that the question generator can generate all specified types of questions; the Stuff Ontology [12], developed by the same author as AWO, which is a core ontology about categories of 'stuff', such as pure and mixed stuff, colloids (e.g., foam, emulsion) and solutions; and the BioTop [3] top-domain ontology that provides definitions for the foundational entities of biomedicine. For the first 2 experiments, we only used the extended version of AWO, while all three were used for the third experiment.

Methods The methods of the three evaluations are as follows. *First experiment:* Each participant ($n = 5$) evaluated 30 questions generated by Appr 1 using AWO and the templates with DOLCE categories (Type B). Students from the University of Cape Town (UCT) were recruited to complete the evaluation. All participants have at least a secondary school pass of English and can speak English fluently. *Second experiment:* Each participant ($n = 6$) evaluated 40 questions generated by Appr 2, using AWO and the subject domain-tailored templates (Type C). The requirements for each participant are the same as for the first experiment. Each evaluator could participate in either the first or second experiment or in both. We used a pass/fail mechanism for both evaluations to determine whether a sentence conforms to English syntax and semantics. All evaluators were allowed to comment on each sentence and encouraged to do so if the answer was negative. *Third experiment:* 95 questions were generated from the three ontologies using Appr 3. From an ontology, for each type of questions, axioms satisfying the axiom prerequisites are randomly selected for the question generation using the Type D templates. We generated 39, 12 and 44 questions from AWO, Stuff Ontology and Biotop Ontology, respectively. The difference is due to having more or less content satisfying the prerequisites. The 12 questions from the Stuff Ontology still do cover all groups of questions. Seven students and one staff member at UCT ($n = 8$) who have English as their first language or speak English fluently (self-declaration) participated in the evaluation. Only two of them participated in the first two experiments. Each participant evaluated all 95 generated questions and had to answer whether each question is syntactically and semantically correct, choosing between A: Very Good, B: Good, C: Average, D: Bad, and E: Very Bad. Their differences were explained to the participants during the meeting before evaluating the generated questions. All evaluators were allowed to comment on each sentence. We use the central tendency (the median for ordinal values) to determine the quality of the questions.

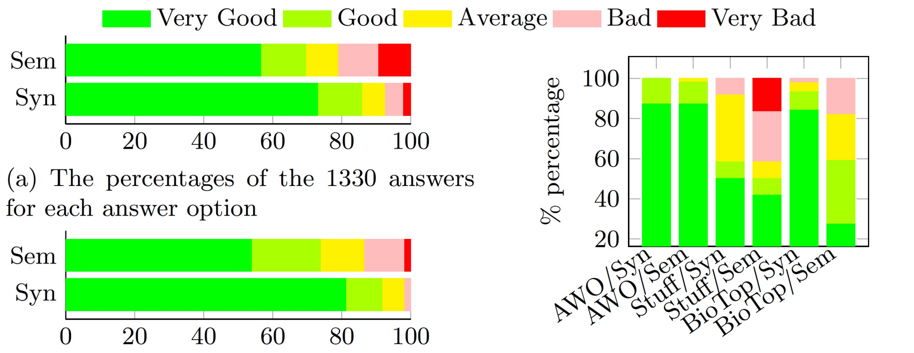
4.2 Results

Appr 1 with Type A and B templates generated some correct questions, such as "Does a herbivore walk?", but the majority failed semantically or syntactically, such as "Is the fly eaten by the walk?". Overall, 26% of the generated questions were considered as quality questions.

Appr 2 with subject domain-specific (Type C) templates generated some correct questions such as: "Does a carnivore eat a terrestrial?" and "True or False: A warthog eats some omnivore.", but also semantically nonsensical ones, such as "Did the terrestrial participate in all the hibernate?". Overall, 34% of the generated questions were considered as quality questions.

For the third experiment, with Appr 3, some of the good generated questions are: "Does a bumble bee fly?" and "True or false: A collective process has a granular part that is a process.". Of the ones classified as 'bad' by the participants, some indeed are, yet others as not (discussed below); questions include "Does a mixed stuff have a part stuff that is a stuff?". In analysing the data, it was evident that one of the eight participants did not perform a proper assessment but randomly selected answers since some good questions were evaluated as bad and vv.; e.g., "Does a carnivorous plant eat an animal?" was labelled with 'Very Bad' and "A condition is a condition of only a situation. True or false?" as 'Good', which is not the case. Therefore, we chose not to consider this participant in further analysis.

The seven participants gave feedback on a total of 665 sentences for syntactic and semantic evaluation; hence, we have 1330 data points. Figure 1a shows the percentage of answers from the evaluators for each answer option (Very Good, $\cdots$, Very Bad), and Figure 1b presents, in percentage, the quality of the generated questions, which refers to the median of the set of evaluations of each question. For the syntax (Figure 1b), 81.05% of the generated questions were classified



(a) The percentages of the 1330 answers for each answer option

(b) The percentages of the 95 questions that were classified as V Good, $\cdots$, V Bad

(c) Syn and Sem quality of the generated questions for each ontology

Fig. 1: Aggregate results of the human evaluation; Syn: syntax; Sem: Semantics.

'Very Good' (77 out of 95 questions); hence, given the ordinal values ordering and the number of participants, at least four evaluators judged the syntax of the question as 'Very Good'. Regarding semantics (Figure 1b), 53.68% and 20% of the questions were 'Very Good' and 'Good', respectively, based on their central tendency. Disaggregating by ontology, the results are as shown in Figure 1c, from which it can be noted that the results for AWO are better than those from the others. We confirmed with a statistical hypothesis test (Fisher's exact) that the results are statistically significantly different (p-value=0.003887 for the syntax and p-value = 3.733e-08 for the semantics). Regarding the inter-rater agreement, the Fleiss Kappa coefficients computed with R language are $k = 0.0856 > 0$ and $k = 0.11 > 0$ for the syntax and the semantics, respectively, which both mean 'slight agreement'. Then, overall, 4 out of 7 evaluators agreed on a single assessment on 85.26% and 60% of the questions generated for their syntax and their semantics, respectively.

4.3 Discussion

From the first two experiments, one can see that specifying the template to a lower level class token helps improve the quality of the generated questions. However, the need for tailoring the generic templates to a specific domain ontology increases the manual effort and decreases the generalisability of question generation across domains. Analysis of the feedback provided by the participants from the first two experiments gave insights as to why the quality rate of generated questions was so low, which amount to two major issues causing the low quality:

- the use of domain and range of OPs to generate questions since it could select unconnected classes, and
- slots that do not adapt to the names of the ontology elements inserted, or: there is a large variation in naming elements within and across ontologies that a fixed template cannot cater for. Since the approach would ideally work for a range of ontologies, it suggested that a reverse order—find the right template for a given axiom—may be a better strategy.

The analysis of the first two approaches assisted in designing Appr 3 and to focus on linguistic aspects instead. This had a much larger improvement in question quality compared to tailoring a generic template to a domain ontology.

The 'slight agreement' between evaluators may come from their different levels of strictness, the disagreement on the place of the word "only" in the questions and the difficulty to understand difficult questions from specific domains, especially for those from Stuff and BioTop Ontology. Understandability of educational questions also straddles into educational research and language proficiency, which is beyond the current scope.

Challenges for generating and evaluating questions There are three main persistent challenges, which affect either the quality of the questions or the user's perception thereof. First, there are words with more than one POS category that are hard to disambiguate in the ontology cf. within-sentence disambiguation for POS tagging; e.g., 'stuff' that can be a noun or a verb.

Second, there is the 'hasX' naming issue of OPs, such as hasTopping, that have already the name of the range in its name. This then results in generated questions such as "Which condition has a life that is some life?", but that ideally would end up as "Which condition has a life?". Furthermore, the question "Does a mixed stuff have a part stuff that is a stuff?" is correct but not 'nice', because the word 'stuff' is repeated 3 times due to the ontology elements MixedStuff, Stuff, and hasStuffPart. Ideally, it would recognise such corner cases and render the question as "Does a mixed stuff have a part that is also a stuff?". Refinement could be made to Algorithm 2 to accommodate for this style of naming OPs after determining several possible OP naming variants, though the algorithm likely always will run behind a modeller's varied naming practice.

Third, there are misunderstood questions, which is an issue that is also unlikely ever to be resolved. The AWO contains general knowledge and is easy for most people to understand. However, since the Stuff and BioTop ontologies are in a specialised domain, we obtained 'Bad' evaluations for some generated questions. For instance, "A mixed stuff has a part stuff that is a stuff. True or false?" is syntactically and semantically correct but was misunderstood by most participants. Also, words with a specific ontological meaning, such as inhere in, were not appropriately assessed; e.g., "True or false: A process quality inheres in a process." is correct but was misunderstood.

Generalisability of education question generation As can be observed in Figure 1, the questions generated from the AWO were evaluated as better than those from Stuff and BioTop. There are three possible reasons for this: either template overfitting, or the AWO additions for coverage testing, or because it was common-sense knowledge, cf. specialised domain knowledge. As stated above, some questions from Stuff and BioTop were misunderstood during the evaluation. In addition, AWO does not have OPs with the "hasX" naming scheme, while the two other ontologies do. Finally, for Stuff Ontology, the word "stuff" has several POS tags, and this affects the quality of the generated questions.

Even though BioTop was not developed by the same developer as the AWO and Stuff ontologies, one can see that the results from BioTop are better than those from the Stuff Ontology. So, this may suggest that Appr 3 with Type D templates and Algorithm 2 has not been overfitted to the modelling style of the AWO developer, therewith indicating potential for generalisability.

Furthermore, we commenced with assessing potential usefulness of our approach for preserving cultural heritage. As a first step in this direction, we generated 3632 questions by using Appr 3 with Algorithm 2 from 3 DH ontologies: Cultural-ON [11] (306 questions), Copyright Ontology (280) and Latin Dance Ontology (3046). For instance, if one can link a dance textbook annotated with the Latin Dance Ontology, one can reuse those generated questions to develop an educational game. Those details and generated questions are available from the supplementary material for further analysis and use. A cursory evaluation indicates that, although our algorithm does not yet cover all corner cases of the myriad of vocabulary naming practices used in ontologies and similar artefacts, there are relevant and good educational questions, such as "What is cross body lead a part of?" and "Does a catalogue describe a collection?".

Overall, it can be concluded that Appr 3 with Type D templates and Algorithm 2 results in good quality questions and generalisability, answering Question 2 from the Introduction (Section 1) in the positive.

5 Conclusions

Three approaches to answerable question generation from ontologies were proposed, involving the specification of axiom prerequisites, a foundational ontology, NLP techniques, template design, and the design and implementation of their respective algorithms. The human evaluation showed that the NLP-based approach (Appr 3 with Type D templates and Algorithm 2) outperformed the others by a large margin. The generated questions from 3 ontologies in different domains were deemed for 80% to have very good syntactic quality and 73.7% very good or good semantic quality. The results also indicated good prospects of generalisability of the proposed solution to ontologies in other subject domains.

Current and future work involves various extensions, including improving on the questions generated from the DH ontologies, more combinations of prerequisites to generate educationally more advanced questions, and link them to annotated textbook text.

Acknowledgements TR acknowledges support from the Hasso Plattner Institute for Digital Engineering through the HPI Research School at UCT.

References

1. Abacha, A.B., Dos Reis, J.C., Mrabet, Y., Pruski, C., Da Silveira, M.: Towards natural language question generation for the validation of ontologies and mappings. Journal of Biomedical Semantics **7**(1), 1–15 (2016)
2. Alsubait, T., Parsia, B., Sattler, U.: Ontology-based multiple choice question generation. KI - Künstliche Intelligenz **30**(2), 183–188 (Jun 2016)

3. Beisswanger, E., Schulz, S., Stenzhorn, H., Hahn, U.: BioTop: An upper domain ontology for the life sciences. Applied Ontology **3**(4), 205–212 (2008)
4. Bouayad-Agha, N., Casamayor, G., Wanner, L.: Natural language generation in the context of the semantic web. Semantic Web **5**(6), 493–513 (2014)
5. Bühmann, L., Usbeck, R., Ngomo, A.C.N.: Assess—automatic self-assessment using linked data. In: Proc. of ISWC. pp. 76–89. Springer (2015)
6. Chaudhri, V.K., Clark, P.E., Overholtzer, A., Spaulding, A.: Question generation from a knowledge base. In: Proc. of EKAW'14. pp. 54–65. Springer, Cham (2014)
7. EV, V., Kumar, P.S.: Automated generation of assessment tests from domain ontologies. Semantic Web **8**(6), 1023–1047 (2017)
8. Gatt, A., Reiter, E.: SimpleNLG: A realisation engine for practical applications. In: Proc. of ENLG'09. pp. 90–93 (2009)
9. Graesser, A.C., Person, N., Huber, J.: Mechanisms that generate questions. Questions and information systems **2**, 167–187 (1992)
10. Hovy, E., Gerber, L., Hermjakob, U., Junk, M., Lin, C.Y.: Question answering in webclopedia. In: Proceedings of the 9th Text retrieval conference (TREC-9) (2001)
11. Italian Ministry of Cultural Heritage and Activities: Italian institute of cognitive sciences and technologies, cultural-on (cultural ontology): Cultural institute/site and cultural event ontology, link: `http://dati.beniculturali.it/cis/3.2` (2016)
12. Keet, C.M.: A core ontology of macroscopic stuff. In: Proc. of EKAW'14. LNAI, vol. 8876, pp. 209–224. Springer (2014), 24-28 Nov, 2014, Linkoping, Sweden
13. Keet, C.M.: The african wildlife ontology tutorial ontologies. J Biomed Semant **11**(4) (2020)
14. Khodeir, N.A., Elazhary, H., Wanas, N.: Generating story problems via controlled parameters in a web-based intelligent tutoring system. The International Journal of Information and Learning Technology **35**(3), 199–216 (2018)
15. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: Ontology library. WonderWeb Deliverable D18 (ver. 1.0, 31-12-2003). (2003), http://wonderweb.semanticweb.org
16. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
17. Ngomo, A.C.N., Moussallem, D., Bühmann, L.: A holistic natural language generation framework for the semantic web. arXiv preprint arXiv:1911.01248 (2019)
18. Olney, A.M., Graesser, A.C., Person, N.K.: Question generation from concept maps. Dialogue & Discourse **3**(2), 75–99 (2012)
19. Papasalouros, A., Kanaris, K., Kotis, K.: Automatic generation of multiple choice questions from domain ontologies. In: Proc. of IADIS International Conference on e-learning. pp. 427–434 (2008)
20. Rodríguez Rocha, O., Faron Zucker, C.: Automatic generation of educational quizzes from domain ontologies. In: Proc. of EDULEARN. pp. 4024–4030 (2017)
21. Rodríguez Rocha, O., Faron Zucker, C.: Automatic generation of quizzes from dbpedia according to educational standards. In: The Third Educational Knowledge Management Workshop. pp. 1035–1041 (2018), lyon, France. April 23 - 27, 2018
22. Sirithumgul, P., Prasertsilp, P., Suksa-ngiam, W., Olfman, L.: An ontology-based framework as a foundation of an information system for generating multiple-choice questions. In: Proc. of the 25th AMCIS (2019)
23. Vinu, E.V., Sreenivasa Kumar, P.: A novel approach to generate mcqs from domain ontology: considering dl semantics and open-world assumption. Journal of Web Semantics **34**, 40–54 (2015)
24. Zhang, L., VanLehn, K.: How do machine-generated questions compare to human-generated questions? Research and practice in technology enhanced learning **11**(1), 7 (2016)