Design and evaluation of an ontology-based tool for generating multiple-choice questions

Marija Cubric and Milorad Tosic

Abstract

Purpose – The recent rise in online knowledge repositories and use of formalism for structuring knowledge, such as ontologies, has provided necessary conditions for the emergence of tools for generating knowledge assessment. These tools can be used in a context of interactive computer-assisted assessment (CAA) to provide a cost-effective solution for prompt feedback and increased learner's engagement. The purpose of this paper is to describe and evaluate a tool developed by the authors, which generates test questions from an arbitrary domain ontology, based on sound pedagogical principles encapsulated in Bloom's taxonomy.

Design/methodology/approach – This paper uses design science as a framework for presenting the research. A total of 5,230 questions were generated from 90 different ontologies and 81 randomly selected questions were evaluated by 8 CAA experts. Data were analysed using descriptive statistics and Kruskal– Wallis test for non-parametric analysis of variance.

Findings – In total, 69 per cent of generated questions were found to be useable for tests and 33 per cent to be of medium to high difficulty. Significant differences in quality of generated questions were found across different ontologies, strategies for generating distractors and Bloom's question levels: the questions testing application of knowledge and the questions using semantic strategies were perceived to be of the highest quality.

Originality/value – The paper extends the current work in the area of automated test generation in three important directions: it introduces an open-source, web-based tool available to other researchers for experimentation purposes; it recommends practical guidelines for development of similar tools; and it proposes a set of criteria and standard format for future evaluation of similar systems.

Keywords Computer-assistedassessment,Design-scienceresearch,Multiple-choicequestion, Ontologies, Automatic question generation

1. Introduction

Objective tests comprise a set of questions with predetermined correct answers. For instance, such a question might ask a user to select an answer from a set of choices, as in a multiple-choice question, to fill in the blank spaces in a paragraph, or to provide a short numeric or textual response. The finite, deterministic nature of the objective tests makes them a good candidate for computer-assisted assessment (CAA), and enables many important elements of a good teaching practice such as encouraging active learning, emphasising time on task, respecting diverse ways of learning and above all, providing immediate feedback to the learner (Chickering and Gamson, 1987).

The impact of assessment feedback on quality and effectiveness of learning has been one of the major topics in education research in the last two decades. The evidence from the research suggests that prompt feedback has significant positive influence on learner’s overall experience, including the engagement, the motivation, the quality of learning and ultimately the improved learning performance (Draper 2009; Govindasamy, 2001; Nicol, 2007). Outside the educational sector, diagnostic tests have been used by business and government’s certification and accreditation bodies. Some of the large-scale CAA technology providers supply test platforms to millions of enterprise users worldwide.

Despite the benefits, the issues such as the availability of reusable question banks (Littlejohn, 2003) and the time required to write good questions are preventing a wider adoption of this form of assessment (Sidick, Barrett and Doverspike, 1994). According to Collins (2006) professional test-writers spend one hour in average per test item. The most popular form of objective test items, the multiple-choice question or MCQ consists of the question text (the stem), one correct answer (the key) and a small number of incorrect options (the distractors). Unlike single-response questions, multiple-response questions include more than one key. The most time-consuming part in an MCQ construction is selecting plausible distractors which could distinguish between different levels of knowledge (Mitkov et al., 2009). The automated and semi-automated means for generating distractors offer a possible solution for this problem.

The critics of objective tests (Biggs, 1999; Paxton, 2001; McKenna, 2019) highlight the issues such as overemphasis on factual knowledge, question design bias and guessing. These issues could start to be addressed by introducing the tools for generating questions which will incorporate specific pedagogical guidelines for assessing different knowledge levels while reducing the (human) bias in the process of the test design. Despite the issues, the critics agree that well-designed MCQs can be very useful in formative assessment “as a means of providing interactive exploration and instant feedback” (McKenna, 2019).

Generating questions from text corpora has been a subject of research and experimentation in areas such as natural language processing (NLP) and pattern-matching. Mitkov et al. (2009) point to the computational difficulties related to the NLP techniques, and conclude that these processing-heavy algorithms, which are based either on the grammar-matching systems or statistically adequate web searches have not yet produced usable systems. The main issue is that these techniques are lacking the “semantic” information for generating plausible distractors.

A recent proliferation of internet-based knowledge repositories, such as Wikipedia, WordNet, Google Knowledge Graphs and others, has created conditions for development of tools and techniques for computer-generated questions based on the semantics of the underlying concepts. This is a result of the availability of the meta-language used for structuring the knowledge in these repositories. For example, Wikipedia’s categories, disambiguation links, headings such as “See also”, and more specific headings for the articles from the same categories, are providing necessary semantic information to enable programmable searches and matches.

Unlike Wikipedia’s semi- and loosely-structured knowledge, ontologies provide a more formal means for representing descriptive, non-procedural knowledge based on a simple meta-language borrowed from computer science. The knowledge in ontologies is organised in a hierarchy of classes, individuals and their properties, which are used to denote the knowledge domain concepts, their instances and their properties respectively. Since the inception of ontologies as the main format for representing knowledge in the semantic web (Berners-Lee, Hendler, and Lassila, 2001) many domain experts and organisations have engaged in their development, particularly in biomedical sciences (see for example, https://bioportal.bioontology.org/)

but also in other fields such as business, linguistic and engineering. In education, ontologies have been used to formalise instructional processes and learning designs (Knight, Gaševic & Richards, 2006), and to support classification of learning materials (Valaski, Reinehr, and Malucelli, 2017). The role of ontologies in designing learning assessment has been less studied and only recently are techniques for ontology-based assessment starting to emerge. In the follow-up section we provide a review of the main work in this area.

The main application area of this paper is automatic generation of MCQ tests from arbitrary domain ontologies.
The paper uses Design Science Research (DSR) approach to extends the current research in this area in three important directions: (i) it introduces an open-source, web-based experimental tool for generating questions from ontologies and from question stem templates which are based on Bloom's taxonomy (1956) ; (ii) it recommends practical guidelines for development of similar tools, based on the evaluations of generated questions (iii) it compares the tool with the similar systems and proposes a set of criteria for future evaluation of test generation tools.

The use of (the original) Bloom's taxonomy as opposed to other alternative learning taxonomies such as Gagne (1965) Anderson et al. (2001), or Marzano (2001), to mention a few, is justified as follows: (i) it is widely used by MCQ developers to develop questions that test knowledge recall, comprehension, application, analysis, synthesis and evaluation (Conole and Warburton, 2005; Nicol, 2007); (ii) it is simple, easy to understand, and it offers a variety of synonyms useful for creating question templates for different levels of the taxonomy (Krathwohl, 2002). The revised Bloom's taxonomy (Anderson et al. 2001) overcomes the limitations of the original taxonomy through: (i) separating knowledge and cognitive process dimensions, (ii) relaxing a strict hierarchy requirement by allowing overlapping levels and (iii) making a more explicit distinction between the synthesis (creation) and the evaluation levels (Krathwohl, 2002). However, none of these improvements are directly relevant to the objectives of this paper i.e. (i) the type of knowledge that can be encoded in ontologies is conceptual knowledge (ii) the tool does not impose a strictly linear progression through the levels of the learning taxonomy, and (iii) the question templates do not include synthesis and evaluation levels. Most importantly, the underlying pedagogical framework embedded in the question templates can be changed with the introduction of different templates, thus allowing other learning taxonomies to be used.

The main contributions that the paper makes to the practice is that it enables a rapid production of questions of different difficulties, that can be used in a context of intelligent tutoring systems, virtual-learning environments, MOOCs and large-scale assessment systems.

The paper is organised according to Gregor and Hevner's (2013) recommendations for presenting the DSR research, and

2. Related Work

In this section, a historical overview of the related work is presented, including a comparison of experimental systems and techniques for computer-generated objective tests (Table 1).

Table 1 .Comparison of experimental systems for generating questions

System (reference)	Type of questions	Knowledge representation	Strategies for distractors	Evaluation method	Learning theory
Fischer (2001)	part-of, application-of	Concept hierarchy	NA	Anecdotal evidence	None
Holohan et al. (2005)	is-a, is-not-a, example-of	OWL ontology	Concept distance	NA	None
Mitkov et al. (2006, 2009)	find-subject, find-object, kind-of	Text and WordNet	Hypernyms & coordinates from WordNet	Expert opinion & CTT	None
*Papasalouros, Kotis, & Kanaris (2008)	choose the correct sentence	OWL ontology with inferred items	11 ontology-based strategies	Expert opinion	None
Authors (2009)	what-is, Is-a, is-not-a	RDF ontology	Same as (*)	NA	None
Authors (2011)	what-is-definition-of, what-is-defined-by, example-of, analogy, what-is-a-generalisation-of, fill-in-the-blank	RDF ontology	Added annotations-based strategies to (*)	NA	Bloom and Krathwohl (1956)
Alsubait, Parsia, and Sattler (2012)	analogy	OWL DL ontology	Semantic similarity: class	Solver method	None
Al-Yahya (2014)	kind-of, fill-in-the-blank	RDF ontology	Random	Expert opinion	None
Vinu and Kumar (2015)	choose-correct answer, analogy	OWL ontology with restrictions	Semantic similarity	Expert opinion	None
Demaidi, et al. (2017)	is-sub-concept-of, example-of, analogy, fill-in-the-blank	OWL ontology	Same as the authors [9]	CTT & IRT	Bloom and Krathwohl (1956)

Fischer (2001) was the first author to explore the use of concept hierarchies in generating test questions. He examined the ways in automating question generation for a technology course within an adaptive hypermedia learning environment (Multibook). The questions were based on “part-of” and “application-of” relations between the subject domain concepts and the distractors were extracted from the “sibling” concepts. The experimental validation of the results showed that in most cases the students were not able to distinguish between the question generated manually and by the system, but whenever more than one super-concept relation was traversed to generate distractors i.e. when the semantic similarity between correct and wrong answers was decreased, the students were able to identify the difference. The idea of using the distance between the nodes in the knowledge graph (the semantic similarity) to model the difficulty of distractors has been used extensively in the subsequent research in this area.

The OntAWare adaptive learning tool developed by Holohan et al. (2005) supported generation of simple learning objects such as slide shows and objective tests using related and unrelated concepts from an input ontology. The proximity of the concepts in the hierarchy was again proposed for generating challenging questions. OntoAWare focused on adaptivity and personalisation, and less so on the implementation of the question generator. Their work was further extended in 2006 to include dynamic problem generation from the relational databases and in from of database queries.

Mitkov, Le, and Karamanis (2006) used NLP techniques combined with WordNet definitions for generating multiple-choice tests, and then applied manual post-editing to improve the quality of the questions. Their starting premises were that the questions should focus on the key concepts, and that the distractors should be as semantically close as possible to the answer. The latter they named as "the distractor selection premise". Their system identified important terms in the text and transformed the comprising sentences into question stems using three rules: find subject, find type ("kind-of" question), and find object. They then retrieved the hyponyms of the correct answer from WordNet to construct the distractors. They were the first authors to perform a comprehensive evaluation of the generated questions (for details see Table 7) which showed that the semi-computerised construction of questions is at least three times more efficient than purely manual production. They also validated the "distractor selection premise" using methods from the Classical Test Theory (CTT). The item difficulty, measured as a percentage of students who answered the question correctly, was lower than for the manually generated tests, while the discriminatory power was higher. In the follow-up work (2009) they performed an empirical evaluation of various similarity measures for generating distractors, but did not find any statistically significant differences.

The work presented here builds on the work of Papasalouros, Kotis, and Kanaris (2008). Their focus was on distractor generation, given a correct answer from an arbitrary domain ontology. They refined the "distractor selection premise" by introducing eleven ontology-based strategies for constructing distractors. The strategies were based on the basic meta-ontology relations between classes, individuals and properties. For example, if the correct answer is "a is an instance of A", one strategy was to offer the instance b of the class B as a distractor, where B was a specialisation (hyponym, "subclass" or "kind") of A. They did not consider stem generation, and used a generic text ("Choose the correct sentence:") instead. After evaluating a selection of questions generated from five purposefully developed ontologies, they concluded that the input ontologies should adhere to certain conventions to generate syntactically correct questions; for example, property name must be written as a verb or a verb-like phrase. They also found that the property-based strategies, compared to class- and relation-based, may produce more but less syntactically correct questions. The main limitations of their work are related to the lack of pedagogical foundation and the use of purposely-constructed ontology for the evaluation. They later (2011) extended the work to include multimedia ontologies.

In the previous work the authors (2009) optimised the strategies for generating distractors introduced by Papasalouros, Kotis, & Kanaris (2008) and implemented a test generator as a plugin for the Protégé ontology editor, using simple question types such as "what-is", "is-a", "is-not-a"; they also introduced the question difficulty as a configurable parameter which was implemented according to the "distractor selection premise". That work was extended in 2011, by adding new strategies that made use of ontology annotations, and by introducing templates for generating questions based on Bloom's taxonomy of learning objectives (Bloom and Krathwohl, 1956).

Alsubait, Parsia, and Sattler (2012) focused on generating analogy questions where the stem is a pair of ontology concepts, and the goal is to identify the most analogous pair of words amongst the given options. Their algorithm was based on the "distractor selection premise" and they evaluated it using an automatic "solver" method. However, the choice of input ontologies did not allow for a realistic assessment as the ontologies were either very complex (Gene ontology) or very simple (Pizza, People and Pets).

Al-Yahya (2014) considered simple questions of "kind-of" and "fill-in-the-blank" types, but with no specific strategies for generating distractors. The system was tested on randomly selected questions from two ontologies. The results of the evaluation indicated that the main problem was the quality of distractors.

Vinu and Kumar (2015) made use of axiomatised knowledge on the concepts from an ontology. They introduced a measure for question difficulty based on the "distractor selection premise". For the experimentation purposes, they considered four ontologies and restricted the evaluation to two question types: "Choose correct answer" and analogy questions. They found that the difficulty of generated questions judged by the experts did not always coincide with the proposed difficulty. The difference was due to the clues in the correct answers. The problem of generating syntactically correct question items was only partially tackled.

Most recently Demaidi, Gaber, and Filer (2017) used the strategies developed by the authors (2011) to evaluate the generated questions using the statistical analysis of the test performance results; they used the methods provided by the CTT and Item Response Theory (IRT), to analyse the results of three tests. Unlike in previous approaches, a sample of questions were selected and syntactically checked before the evaluation. Their results showed that the test could effectively discriminate between high and low performing students. They also showed that the strategies for generating discriminators and the Bloom's taxonomy level (1956) both affect the question difficulty and the discriminatory power: the questions at the knowledge level were easier than the other types, but no significant difference in difficulty was found in the questions testing application, comprehension and analysis. They did not consider the syntactic correctness, as all questions included in the evaluation were previously edited.

While all the reviewed approaches have some important novel elements, only a few provide empirical evidence on the question utility and the difficulty, and none on the relevance and textual quality of the questions. In this paper, we address this gap by providing an evaluation of generated questions according to all four criteria, and comparing the evaluations with those of other related systems discussed in Section 2. To simplify the presentation, we use the abbreviation "Onto2MCQ" to denote a class of systems or approaches for generating MCQs from arbitrary domain ontologies. Following the guidance by Thuan et al. (2019) for construction of design science research questions, we formulate the following research questions that this paper will try to answer:

RQ1. Can an Onto2MCQ be implemented that uses a combination of syntactic and semantic question generation strategies and question templates based on a specific learning taxonomy?

RQ2. What evaluation measures can be used to assess Onto2MCQs ?

RQ3. In what ways can Onto2MCQ be improved?

3. Methodology

Design Science Research (DSR) is a methodology commonly employed in Information System research and other design-oriented disciplines. The methodology involves construction of socio-technical artefacts, such as software, algorithms, procedures, principles or methods, and extraction of the knowledge resulting from the novelty of the artefact or its development process. So, in DSR design becomes a research method for the knowledge construction.

The method follows the stages of a standard design process: (i) awareness of the problem (ii) suggestions for the problem solution drawn from the existing knowledge base and functional requirements (iii) development of the artefact (iv) evaluation of the artefact, where the last three stages can be iterated. In this process, the knowledge flows in two directions: from the existing knowledge base to the “suggestions”, and from the “development” and “evaluation” to the knowledge base.

This paper uses the “pragmatic-design” variant of the DSR (Gregor and Hevner, 2013) where the emphasis is not on theoretical contribution but on development and evaluation of a technical artefact. The artefact under the constructions is a web-based Onto2MCQ. The research covers the first two levels in the DSR contribution type hierarchy (Gregor and Hevner, 2013), as it not only describes and evaluates a software tool (RQ1&RQ2) , but it also identifies the relevant technological rules and design principles that can be used for similar systems (RQ3).

The evaluation of the artifact was based on a mixed-method comprising quantitative and qualitative data obtained from the CAA subject experts combined with the usage data obtained from the server platform and the evidence of the impact in the field. The quantitative data were analysed using descriptive statistics and non-parametric tests; and the qualitative data were collected through semi-structured interviews with the evaluators who represented the two ends of the evaluation spectrum (positive and negative opinions).

4. The Onto2MCQ tool

The Onto2MCQ experimental tool (http://www.opensemcq.org) was developed as a standalone web application using the open-source and free development tools such as Java, PHP, UserCake, MySQL, Apache Jena, Apache PDFBox. The code for the question generation engine was made available as an open source in the GoogleCode repository. The tool, publicly hosted in one of the authors institutions, provides a fully-functional standalone web-based experimental system for generating MCQs (single-response format) from ontologies which (i) uses a combination of semantics and annotations-based strategies for generating distractors (ii) employs templates for generating questions stems corresponding to different knowledge levels in Bloom's taxonomy (1956), including factual knowledge, comprehensions, application and analysis.

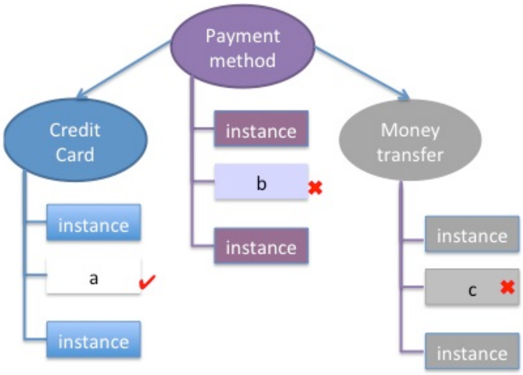
The inclusion of annotations i.e. textual descriptions of the ontology concepts, individuals, and properties to the algorithm was justified by an increasing trend of adding annotations to the ontologies by automatic or semi-automatic means (Reeve and Han, 2005); this has resulted in more ontologies being equipped with textual description. The annotations have enabled new question templates to be used in the process of test generation (e.g. “What is the paragraph describing …?”, “Fill-in the missing words in the text …”, “What is the correct definition of …?”) and consequently, production of more diverse tests.

Other question templates used for generating tests were linked to the strategies previously defined in Papasalouros, Kotis, & Kanaris (2008), and the authors (2011) as illustrated in the example in Figure 1. In the example, the distractors are selected from the instances of the “parent” class (Payment method) and the “sibling” class (MoneyTransfer).

Figure 1. An example question generated from an e-commerce ontology and a model of the strategy for generating the distractors

Which one of the following examples demonstrates the concept “credit card”?

a) American Express
b) *Cash*
c) Direct debit
c) Pay Swarm
d) Google Checkout



A total of seven strategies for generating distractors were implemented:

1) StrConceptToDefinitionJaro (knowledge)
2) StrDefinitionToConceptJaro (knowledge)
3) StrExistingNotExistingRelation (understanding)
4) StrInstanceOfParentClass (application)
5) StrInstanceOfSiblingClass (application)
6) StrParagraphToConceptJaro (analysis)
7) StrDefinitionToSuperConceptJaro (analysis).

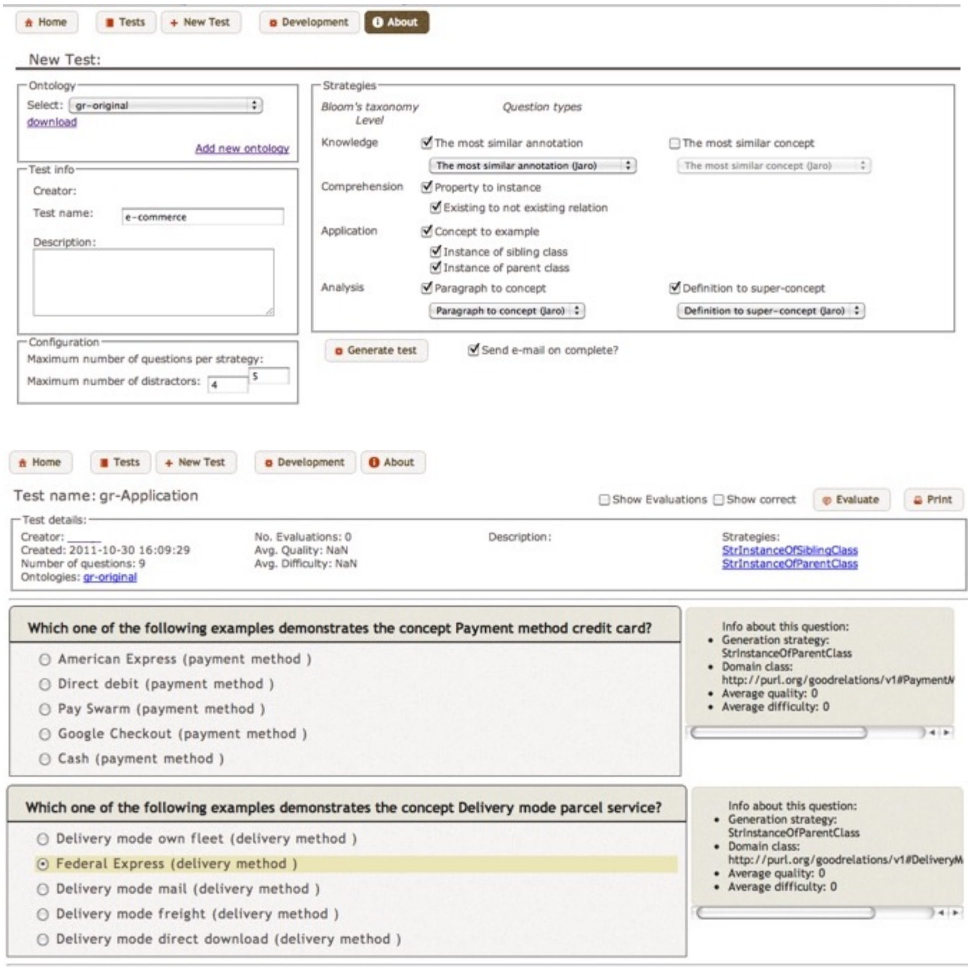
All strategies are based on distractor selection premise and strategies 1-2 and 6-7 also make use of the syntactic similarity of the ontology annotations.

According to the classification from Papasalouros, Kanaris, and Kotis (2008), strategies 3 and 7 are property and sub-concept strategies, while the rest are class-based strategies. The techniques used for generating syntactically similar items include common string-similarity algorithms provided by the open-source Java library SecondString (Cohen, Ravikumar and Fienber, 2003).

Linking the question templates to the Bloom’s knowledge levels (1956) has enabled qualifying the difficulty of the questions according to the increasing educational objectives: factual knowledge (what is the correct definition of…?), understanding (which of the following response pairs relates in the same way as … ?), application (which of the following examples demonstrates …?), analysis (which of the following concepts is a generalisation of …?). Various strategies using distractor-selection-premise were employed for generating questions at different levels.

The screenshots in Figure 2 illustrate the views for creating and viewing the generated tests.

Figure 2. Creating test and viewing test results



New Test:

Ontology
Select: gr-original
download
Add new ontology

Test info
Creator:
Test name: e-commerce
Description:

Configuration
Maximum number of questions per strategy: 5
Maximum number of distractors: 4

Strategies

Bloom's taxonomy Level	*Question types*	
Knowledge	☑ The most similar annotation The most similar annotation (Jaro)	□ The most similar concept The most similar concept (Jaro)
Comprehension	☑ Property to instance ☑ Existing to not existing relation	
Application	☑ Concept to example ☑ Instance of sibling class ☑ Instance of parent class	
Analysis	☑ Paragraph to concept Paragraph to concept (Jaro)	☑ Definition to super-concept Definition to super-concept (Jaro)

☑ Send e-mail on complete?

Test name: gr-Application

□ Show Evaluations □ Show correct

Test details:
Creator:
Created: 2011-10-30 16:09:29
Number of questions: 9
Ontologies: gr-original

No. Evaluations: 0
Avg. Quality: NaN
Avg. Difficulty: NaN

Description:

Strategies:
StrInstanceOfSiblingClass
StrInstanceOfParentClass

Which one of the following examples demonstrates the concept Payment method credit card?

○ American Express (payment method)
○ Direct debit (payment method)
○ Pay Swarm (payment method)
○ Google Checkout (payment method)
○ Cash (payment method)

Info about this question:
- Generation strategy: StrInstanceOfParentClass
- Domain class: http://purl.org/goodrelations/v1#PaymentM
- Average quality: 0
- Average difficulty: 0

Which one of the following examples demonstrates the concept Delivery mode parcel service?

○ Delivery mode own fleet (delivery method)
● Federal Express (delivery method)
○ Delivery mode mail (delivery method)
○ Delivery mode freight (delivery method)
○ Delivery mode direct download (delivery method)

Info about this question:
- Generation strategy: StrInstanceOfParentClass
- Domain class: http://purl.org/goodrelations/v1#DeliveryM
- Average quality: 0
- Average difficulty: 0

5. Evaluation

The Onto2MCQ tool was evaluated using the criteria such as quality, difficulty, validity and utility of the generated questions.
In this paper, the focus is on the evaluation of the test generation engine, rather than usability of the user interface or efficiency of the tool, which will be considered in the follow-up work.

5.1 Evaluation Criteria

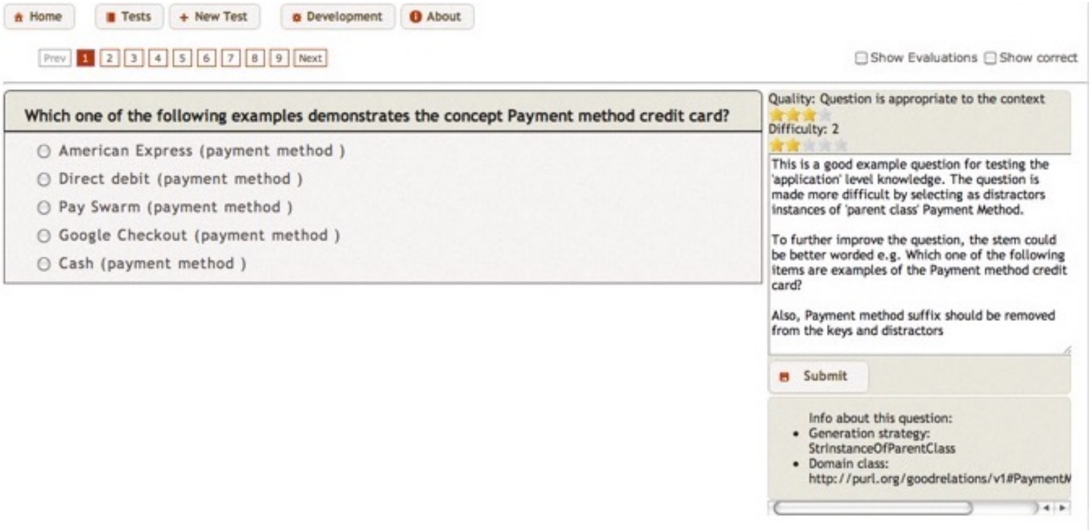
The evaluation criteria were chosen based on the recommendations from Gregor and Hevner (2013): (i) the quality criteria assess the quality of the generated test-items, including the syntactic correctness, the choice of distractors and the knowledge level i.e. the question difficulty; (ii) the validity means that the tool works and does what it is meant to do. In the context of this paper the validity was assessed through the relevancy of the question to the underlying ontology concepts (iii) the utility criteria assess whether the achievement of the goals has value outside of the development environment, i.e. if the question could be used without any changes in a real assessment.

A CAA expert from one of the author's institution reviewed an interim version of the evaluation questionnaire, and further improvements were made based on the recommendations.

The final version of the evaluation questionnaire is included in the Appendix A.

To facilitate the evaluation process, the tool was extended with the evaluation capabilities. Figure 3 shows an example of a generated question and its evaluation. A few more examples of generated questions across various evaluation ratings are shown in Appendix B.

Figure 3. Evaluating questions


Home
Tests
+ New Test
Development
About
Prev 1 2 3 4 5 6 7 8 9 Next
□ Show Evaluations □ Show correct
Submit

Which one of the following examples demonstrates the concept Payment method credit card?

○ American Express (payment method)
○ Direct debit (payment method)
○ Pay Swarm (payment method)
○ Google Checkout (payment method)
○ Cash (payment method)

Quality: Question is appropriate to the context
Difficulty: 2

This is a good example question for testing the 'application' level knowledge. The question is made more difficult by selecting as distractors instances of 'parent class' Payment Method.

To further improve the question, the stem could be better worded e.g. Which one of the following items are examples of the Payment method credit card?

Also, Payment method suffix should be removed from the keys and distractors

Info about this question:

- Generation strategy: StrInstanceOfParentClass
- Domain class: http://purl.org/goodrelations/v1#PaymentM

5.2 Evaluation Ontologies

The selection of the test ontologies from the initial set of 90 ontologies used for generating tests was based on the following criteria: academic subject knowledge domain, different subject domains covering both STEM and non-STEM disciplines, inclusion of annotations, inclusion of other ontology components (concepts, individuals, properties), non-flat concept hierarchy (at least 3 levels), contains at least 10 concepts.

The first three criteria resulted in 16 candidate ontologies, and amongst them the following four were selected for the evaluation purposes, based on the remaining criteria: Biochemistry (http://ontology.dumontierlab.com/biochemistry-complex),Economy (http://relant.teknowledge.com/DAML/Economy.owl), Law (http://www.estrellaproject.org/lkif-core/), and Music (http://purl.org/ontology/mo/).

Table 2 summarises the characteristics of the ontologies selected for the evaluation.

Table 2 Counts and percentages of ontology components for the evaluation ontologies

Ontology components		Concepts		Individuals		Properties		Annotations		Total	
		Count	%	Count	%	Count	%	Count	%	Count	%
Ontology	Biochemistry	1041	34.7	0	0.0	82	3.0	1879	63.0	3002	100.0
	Economics	341	28.0	482	40.0	61	5.0	313	26.0	1197	100.0
	Law	155	30.0	0	0.0	97	19.0	266	51.0	518	100.0
	Music	142	7.0	87	4.0	387	18.0	1496	71.0	2112	100.0
	Total	1679	25.0	569	8.0	627	9.0	3954	58.0	6829	

5.3 Evaluation Questions

The test generation parameters were configured as follows:

- Maximum number of questions per question type: 10
- Maximum number of distractors: 4
- Type of question in Bloom's taxonomy: knowledge, understanding, application, analysis.
- Strategies for distractors (seven strategies described in Section 4).

One test was generated for each of the four ontologies. The total number of questions available for the evaluation was 81, ranging from 19 (biochemistry) to 27 (economy). Thirty-eight questions (47%) were testing factual knowledge while the remaining 43 (53%) were questions related to the higher educational objectives. As expected, the number of generated questions for "comprehension" and "application" levels were directly proportional to the number of "properties" and "individuals" in the corresponding ontologies. Similarly, and as it was the case in the Music ontology, the lack of associations between the classes and individuals have resulted in no "application" questions been generated. Naturally, the structure of the ontology and the number of the associations between different ontology components has determined the diversity and the size of the generated tests.

Strategies using annotations produced 65% of questions; this is followed by class (72%) sub-class (16%) and property-based strategies (11%).

Regarding the evaluation ontologies, Economy ontology produced more diverse and challenging questions, as 17(63%) of its questions are testing higher educational objectives. Biochemistry and Law ontologies generated less diverse tests (only two types of questions) and less challenging questions: 53% of question in both cases are testing only factual knowledge.

The summary of the counts and percentages of the test questions is shown in Table 3.

Table 3 Counts and (row) percentages of the (generated) test questions

Test		Biochemistry		Economy		Law		Music		
		Count	%	Count	%	Count	%	Count	%	Coun
Bloom's level	Knowledge	10	26.3	10	26.3	8	21.1	10	26.3	38
	Comprehension	0	0.0	6	66.7	0	0.0	3	33.3	9
	Application	0	0.0	6	100.0	0	0.0	0	0.0	6
	Analysis&Synth.	9	32.1	5	17.9	7	25.0	7	25.0	28
	Total	19	23.5	27	33.3	15	18.5	20	24.7	81
Strategy for generating distractors	1	5	27.8	5	27.8	3	16.7	5	27.8	18
	2	5	25.0	5	25.0	5	25.0	5	25.0	20
	3	0	0.0	6	66.7	0	0.0	3	33.3	9
	4	0	0.0	5	100.0	0	0.0	0	0.0	5
	5	0	0.0	1	100.0	0	0.0	0	0.0	1
	6	4	26.7	2	13.3	5	33.3	4	26.7	15
	7	5	38.5	3	23.1	2	15.4	3	23.1	13
	Total	19	23.5	27	33.3	15	18.5	20	24.7	81

5.4 Participants

Following an email invitation to fifteen colleagues from one of the authors' University, eight completed the evaluation. All participants were university lecturers who had expertise and interest in the CAA application area. Each participant was asked to evaluate one test that was the closest to their subject expertise and one additional test outside of their expertise area. All participants completed the evaluation of at least one test and one participant completed two evaluations. Two evaluators completed 34% and 45% of Economy and Music questions respectively while other tests were fully completed.

The academic subject areas of participants included business (2), law (1), IT & computer science (3), and life sciences (2). The age of participants was between 25 and 58, and the male to female ratio was 3:5.

The total number of evaluations was N=138, ranging between 14 (music) to 71 (economy). Fifty-nine evaluations (43%) were related to the questions testing factual knowledge while the remaining 79 (57%) were evaluations related to the higher educational objectives. The summary of the counts and percentages of evaluations is shown in Table 4.

Table 4 Counts and (row) percentages of the evaluations

Test		Biochemistry		Economy		Law		Music	
		Count	%	Count	%	Count	%	Count	%
Bloom's Level	Knowledge	20	33.9	24	40.7	8	13.6	7	11.9
	Comprehension	0	0.0	18	90.0	0	0.0	2	10.0
	Application	0	0.0	16	100.0	0	0.0	0	0.0
	Analysis&Synth	18	41.9	13	30.2	7	16.3	5	11.6
	Total	38	27.5	71	51.4	15	10.9	14	10.1
Strategy for distractors	1	10	32.3	13	41.9	3	9.7	5	16.1
	2	10	35.7	11	39.3	5	17.9	2	7.1
	3	0	0.0	18	90.0	0	0.0	2	10.0
	4	0	0.0	13	100.0	0	0.0	0	0.0
	5	0	0.0	3	100.0	0	0.0	0	0.0
	6	8	40.0	4	20.0	5	25.0	3	15.0
	7	10	43.5	9	39.1	2	8.7	2	8.7
	Total	38	27.5	71	51.4	15	10.9	14	10.1
Evaluators		2	25.0	3	37.5	1	12.5	2	25.0
Discipline	STEM	38	43.7	35	40.2	0	0.0	14	16.1
	nonSTEM	0	0.0	36	70.6	15	29.4	0	0.0
	Total	38	27.5	71	51.4	15	10.9	14	10.1
Gender	Female	38	39.2	35	36.1	15	15.5	9	9.3
	Male	0	0.0	36	87.8	0	0.0	5	12.2
	Total	38	27.5	71	51.4	15	10.9	14	10.1

5.5 Evaluation Results

The frequencies of the responses (0-4) across four evaluation criteria are shown in Table 5. The median values for all four criteria was 1, indicating the following tendencies in data:

- The questions were easy to understand, despite syntactic or grammatical mistakes
- The difficulty of the questions was low i.e. mainly testing factual knowledge
- The questions were not directly useable, but they could serve as a basis for useable question
- The questions relate to some extent to the concepts they are testing.

Table 5 Counts and percentages of the response values; The values for quality (0-2) were mapped to the 0-4 range as follows: 0->0, 1-> 2 and 2->4.

Response value			0	1	2	3	4	Total
Criteria	Quality	Count	63	54	21	0	0	138
		%	45.7	39.1	15.2	0.0	0.0	100.0
	Difficulty	Count	41	65	28	4	0	138
		%	29.7	47.1	20.3	2.9	0.0	100.0
	Utility	Count	43	47	8	27	13	138
		%	31.2	34.1	5.8	19.6	9.4	100.0
	Validity	Count	37	65	27	7	2	138
		%	26.8	47.1	19.6	5.1	1.4	

Table 6 includes the comparison of the medians and the corresponding inter-quartile ranges of the responses across different tests, levels in the Bloom's taxonomy (1956), strategies for generating distractors, subject areas and the gender of the evaluators. Kruskal-Wallis test was used for group comparisons; the test compares the medians of the groups, and it is adequate for comparing more than 2 groups when parametric assumptions are not met. The use of the non-parametric test is justified because the data considered here is not interval data. The resulting p-values are shown in the last row for each of the groups of the evaluations indicating significant differences in quality of questions across different ontologies, question types, strategies and evaluators.

The evaluation was followed-up with unstructured interviews with two evaluators from the law and business subject areas, who assigned the highest and the lowest marks to the questions respectively. The interviews were focused around the topics of usability of the tool in the specific subject areas and the further improvements. The answers revealed the differences in the type of knowledge and assessment in the two domains. While the multiple-choice and concept-oriented questions were commonly used in Law subjects, they are less common in Marketing subjects where "the-generalisation-of" questions where found particularly unsuitable. Improvements were suggested in the areas of clarity of questions, availability of the ontology views and the necessity of providing the explanation on why the answer is wrong.

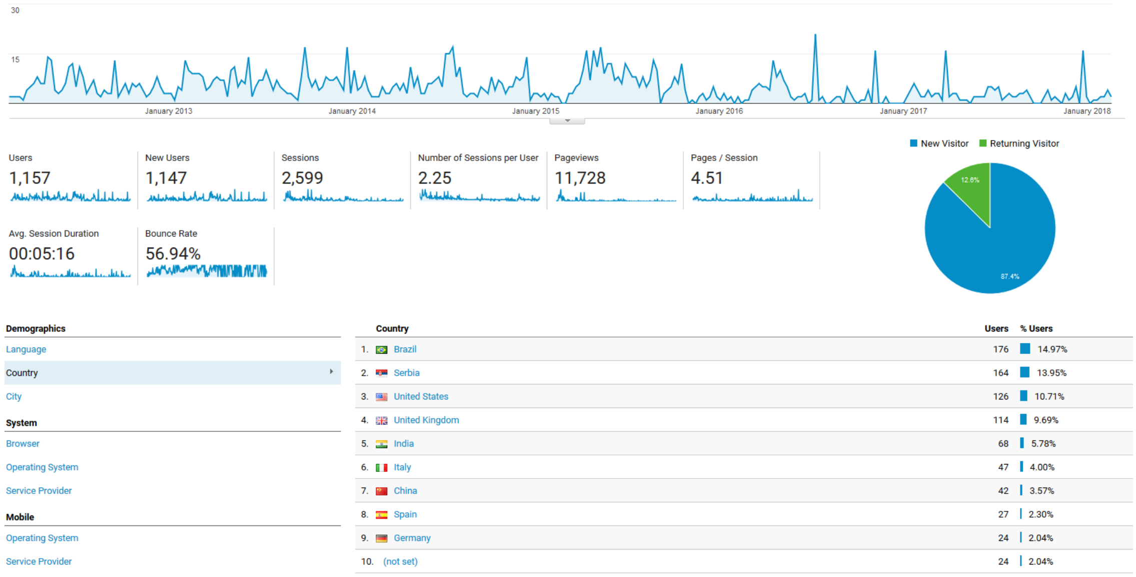
Table 6 Comparison of medians, interquartile ranges (IQR) and p values (*p<0.05, **p<0.01) of the evaluations

Evaluation criteria		Quality	Difficulty	Utility	Validity	Overall Mark
		Median (IQR)	Median (IQR)	Median (IQR)	Median (IQR)	Median (IQR)
Test	Biochemistry (n=38)	2 (2,3)	2 (1,2)	2 (1,4)	2 (2,3)	1.75 (1.25,2.75)
	Economy (n=71)	3 (2,3)	2 (1,3)	2 (2,4)	2 (1,3)	2 (1.5,2.75)
	Law (n=15)	3 (2,4)	2 (1,3)	2 (1,2)	2 (1,2)	2 (1.25,2.5)
	Music (n=14)	2 (2,3)	2 (1,2)	2 (1,4)	2 (1,2)	1.75 (1.5,2.25)
	Total (N=138)	3 (2,3)	2 (1,2)	2 (1,4)	2 (1,3)	2 (1.5,2.5)
	p-value	0.047*	0.464	0.125	0.365	0.49
Bloom's Level	Knowledge (n=59)	3 (2,3)	2 (2,2)	2 (1,4)	2 (1,3)	2 (1.5,2.75)
	Comprehension (n=20)	3 (2,3)	3 (1,3)	2 (2,4)	2 (1,3)	2 (1.5,2.88)
	Application (n=16)	3 (3,4)	2 (1,3)	2 (2,4)	2 (2,3)	2.25 (1.75,2.88)
	Analysis&Synth (n=43)	2 (2,3)	2 (1,2)	2 (1,3)	2 (1,2)	1.75 (1.25,2.25)
	Total (n-138)	3 (2,3)	2 (1,2)	2 (1,4)	2 (1,3)	2 (1.5,2.5)
	p-value	0.046*	0.242	0.142	0.702	0.177
Strategy for Distractors	1 (n=31)	2 (2,3)	2 (1,2)	2 (1,2)	2 (1,2)	1.75 (1.25,2.25)
	2 (n=28)	3 (3,4)	2 (2,2)	3 (2,4)	3 (2,3)	2.5 (1.88,2.88)
	3 (n=20)	3 (2,3)	3 (1,3)	2 (2,4)	2 (1,3)	2 (1.5,2.88)
	4 (n=13)	3 (3,4)	2 (1,3)	2 (2,3)	2 (2,2)	2 (1.5,2.25)
	5 (n=3)	3 (3,3)	2 (2,3)	4 (4,4)	3 (2,3)	2.75 (2.5,3)
	6 (n=20)	2 (2,3)	2 (1,2)	2 (1,4)	2 (1,3)	1.75 (1,2.63)
	7 (n=23)	2 (2,3)	2 (1,2)	2 (1,2)	2 (2,2)	1.75 (1.5,2.25)
	Total (N=138)	3 (2,3)	2 (1,2)	2 (1,4)	2 (1,3)	2 (1.5,2.5)
	p-value	0.019*	0.311	0.06	0.006**	0.022*
Evaluator's Discipline	STEM (n=87)	2 (2,3)	2 (1,2)	2 (1,4)	2 (1,2)	2 (1.5,2.5)
	nonSTEM (n=51)	3 (2,3)	2 (1,3)	2 (1,4)	2 (2,3)	2 (1.5,2.75)
	Total (N=138)	3 (2,3)	2 (1,2)	2 (1,4)	2 (1,3)	2 (1.5,2.5)
	p-value	0.006**	0.199	0.455	0.131	0.299
Evaluator's Gender	Female (n=97)	2 (2,3)	2 (1,2)	2 (1,4)	2 (1,2)	2 (1.5,2.5)
	Male (n=41)	3 (2,3)	2 (1,3)	2 (1,4)	2 (1,3)	2 (1.5,2.75)
	Total (N=138)	3 (2,3)	2 (1,2)	2 (1,4)	2 (1,3)	2 (1.5,2.5)
	p-value	0.005**	0.644	0.923	0.532	0.385

5.5 Impact

Since the early prototype was introduced in 2013 the tool was used by more than 1000 external visitors from more than 10 different countries with the average session duration of 5 minutes and an average bounce rate (Figure 4); during that time, more than 90 different ontologies were uploaded to create 251 tests with 5230 questions. It should be noted that no advertisement has been done except through the published papers and conferences. Hence, the external visitors are probably researchers in the field resulting in a noticeable impact in the research community. The prototype and the method have been referenced in 50 publications, including books, journal and conference papers and the tool has been used for providing experimental support and data for work on 2 ongoing PhD dissertations.

Figure 4. Usage data on Google Analytics



Users	New Users	Sessions	Number of Sessions per User	Pageviews	Pages / Session
1,157	1,147	2,599	2.25	11,728	4.51

Avg. Session Duration	Bounce Rate
00:05:16	56.94%

	Country	Users	% Users
1.	Brazil	176	14.97%
2.	Serbia	164	13.95%
3.	United States	126	10.71%
4.	United Kingdom	114	9.69%
5.	India	68	5.78%
6.	Italy	47	4.00%
7.	China	42	3.57%
8.	Spain	27	2.30%
9.	Germany	24	2.04%
10.	(not set)	24	2.04%

6. Findings and Discussion

In this section, the main findings from the evaluation are discussed, the tool is compared to the similar systems across a range of evaluation criteria and areas of future research are recommended.

The analysis of the results indicates that the textual quality of the generated questions varied across different types of questions, ontologies and strategies. The items testing the application of knowledge through identification of examples of specific concepts are shown to be easier to understand compared to the other question types; the higher-order question types such as analysis questions, were more difficult to understand. This could be attributed to the use of the syntactic similarity strategies for generating distractors in the former case, and to the artificial formulation of the question text in the latter example, which was driven by the ontology meta-language (e.g. the-generalisation-of questions).

The strategies based only on semantic similarity have generated questions that are perceived to be of a higher textual quality. This might be due to less text and thus, less syntactic errors in these questions.

The difficulty of the questions testing factual knowledge, comprehension and application was assessed to be matching the expected knowledge levels, while the analysis questions were evaluated to be easier than expected.

The myth that objective tests cannot test higher orders skills has resulted in a very few examples of these question types in practice. On contrary, the ontology graphs provide means not only for creating more difficult questions but also for generating question templates from the graph sub-components. The challenge is to make those templates accessible using the specific domain language. This approach could be supported with the machine learning techniques for improving the language quality.

Validity of the questions and the overall evaluation marks were significantly different across different strategies. The strategies using only “distractor selection premise” generated questions of higher validity and higher overall value indicating that overall these questions were valued more by the evaluators.

Not all questions were assessed to be directly useable, and many are requiring syntactic improvements. The text included from the ontology includes annotations, as well as the names of the components (classes, individuals and properties) which are frequently using artificial (programming language) syntax and conventions. A pre-processing of the ontology text would therefore be required for any real-life use of the tool.

The question feedback (true/not true) is not always useful and more substantial feedback is required to support the learners and provide opportunities for “deep learning”. With the approach presented in this paper, the feedback is already encoded in the strategy for generating the distractor. Further work is needed in translating these strategies to accessible language.

The quality of the ontologies considered for the evaluation has made an impact on the quality and difficulty of the questions. Therefore, in future work, real life ontologies should be considered as they are more likely to undergo quality assurance. This requires the performance improvements of the tool, or the techniques for splitting the large ontologies into smaller useable components.

Regarding the individual attitudes, the evaluators from the STEM disciplines were more critical of the quality of the questions than their non-STEM colleagues. This could be attributed to the more experience with the multiple-choice questions in the STEM disciplines, and subsequently, higher expectations from the experts from these disciplines.

The extreme difference is evaluations between the law and the business domain experts suggests the differences in the type of knowledge in these domains (rule-based vs. procedural, how-to knowledge) and subsequently different assessment requirements.

The current system is fully automated, and while automation reduces the cost of test development the issues that creates such as some of the questions lacking utility or comprehension, may incur costs that are prohibitive of deployment.

The comparison of the evaluations across different experimental systems, including the authors’ is presented in Table 7, showing improvements in percentage of questions evaluated to be useable after small syntactic corrections (60%) and a number of methodology improvements such as: number of expert evaluators (8), variety of the evaluation criteria (4), total number of generated questions for the evaluation (90), number of ontologies used in the evaluation (4) , richness of the underlying ontologies with regard to the variety of components (concepts, individuals, annotations and properties). In addition to that the results demonstrate significant differences in evaluations across different strategies for generating distractors and different question types confirming similar findings by Demaidi et al. (2017).

Table 7. Comparison of the systems for generating questions; NA = Not Applicable; U= Unknown; *questions edited prior to the evaluation

Authors/systems		Mitkov et al. 2006	Papasalouros et al. 2008	Alsubait et al. 2012	Al-Yahya 2014	Vinu & Kumar, 2015	Demaidi et al. 2017	Author
Method	Evaluation method	expert opinions	expert opinions	solver	expert opinions	expert opinions	CTT & IRT	expert opinior
	#evaluators	2	2	NA	3	7	NA	8
	Evaluation criteria	Syntax, utility	Syntax, utility	Correct %	Syntax, utility	Utility, difficulty	Correct %, difficulty	Syntax difficult utility, validity
Total number of questions generated/ from ontologies		575 /NA	374/5	56/3	1033/2	656/1	NA	5230/9
Evaluation ontology components	ontologies	NA	1	3	2	1	2	4
	concepts	NA	29	58-36146	20-24	105	U	142-10
	individuals	NA	40	NA	73-145	546	U	0-482
	properties	NA	41	NA	25-38	15	U	61-387
	annotations	NA	NA	NA	NA	NA	U	313-18
Evaluation questions	#questions	575	88	56	120	31	44*	81
Evaluations	difficulty - high	NA	NA	NA	NA	16%	7%	4%
	difficulty - medium	NA	NA	NA	NA	23%	48%	29%
	difficulty - low	NA	NA	NA	NA	61%	45%	67%
	Directly useable	3.5%	75%	NA	0%	0%	NA	9%
	Useable w. corrections	53.5%	25%	NA	60-82%	71%	NA	60%
	Unusable	43%	0%	NA	29%	29%	NA	31%
	textual quality (median)	NA	NA	NA	NA	NA	NA	2
	Validity (median)	NA	NA	NA	NA	NA	NA	2
	Item difficulty	75%	NA	8-88%	NA	NA	50%	NA
	Discriminatory power	40%	NA	U	NA	NA	30%	NA
Significant differences across …	strategies	No	U	U	U	U	Yes	Yes
	ontologies	U	U	U	U	U	U	Yes
	question types	U	U	U	U	U	Yes	Yes
	evaluators	U	U	U	U	U	NA	Yes

The table also suggests a standard format (methodology) for reporting future

Onto2MCQ system evaluations to include:

1. Evaluation characteristics:
    - Type of evaluation: expert opinion (include number of evaluators), CTT/IRT, solver, or hybrid
    - Evaluation criteria: quality of text, utility, difficulty, validity, % of correct answers and discriminatory power (for CTT/IRT and solver based methods)
2. Target ontologies:
    - Number of ontologies
    - Number of concepts, annotations, individuals and properties per ontology
3. Questions
    - Type of questions (single-response, multiple-response)
    - Number of discriminators per question
4. Results:
    - Number of generated questions per ontology
    - Number of (un-edited) questions used for the evaluation per ontology
    - Number of (edited) questions used for the evaluation per ontology
    - For all evaluation criteria:
        - Interquartile ranges and median values
        - Significant difference ($p<0.05$) across
            - Strategies used for discriminators
            - Evaluation ontologies
            - Question types
            - Evaluators (if using expert opinion).

The main limitations of the evaluation presented here could be addressed by including a variety of subject experts and other stakeholders (e.g. students and tool developers) in the evaluation and combining the expert-opinion with the test performance analysis based on the CTT and IRT methods.

With regard to the limitations of the questions generator, adding different question types (e.g. multiple-response questions) and adding more question templates (based on different learning taxonomies) are priority areas for future extensions of the system.

A summary of practical guidelines and methods for developing similar systems and suggested areas for future research is provided in Table 8.

Table 8 Recommendations for design and evaluation of test generation systems and areas for future research

R1	A trade-off needs to be considered between (i) using ontology meta-language capabilities for generating questions and (ii) compatibility of the generated question types with the standards in the specific domain.
R2	The "distractor selection premise" could be used to increase difficulty of generated questions at all levels of the taxonomy.
R3	The ontology selected for generating questions should be pre-processed to improve the language where: (i) the text of annotations should be checked for syntactic correctness and (ii) all ontology components should be provided with user-friendly labels.
R4	The strategies for generating distractors could be translated into natural language explanations that can be used for generating question feedback.
R5	A trade-off between generic Ont2MCQ vs. subject-specific Onto2MCQ needs to be considered.
R6	A trade-off between full automation and the cost of language improvements needs to be considered.
R7	The Ont2MCQ evaluation methodology and reporting should be standardised.

7. Conclusions

This paper uses a DSR approach to present and evaluate an experimental web-based Onto2MCQ tool that uses a combination of syntactic and semantic question generation strategies and question templates based on Bloom's taxonomy (RQ1). The Onto2MCQ tool addresses two important problems in the domain of intelligent tutoring systems, namely, providing prompt feedback to the learner, and offering a cost-effective assessment solution for the test provider. The tool was evaluated according to the quality, difficulty, validity and utility of the generated questions and the findings were compared to the related work in this area indicating improvements in many of the evaluation categories (RQ2). The limitations of the tool and of the evaluation method have been recognised; and practical guidelines for improvements and future research areas have been recommended (RQ3).

The recent rise in development of online data vocabularies, increased potential for large-scale experimentation in massive open online courses, and use of machine-learning techniques are opening new possibilities for advances in development of Onto2MCQ systems. Lowering the cost of question creation through increased automation can reduce barriers for a wider adoption of CAA and provide opportunities for higher learner's engagement.

This paper demonstrates that the subject of generating objective tests from ontologies is maturing and that new hybrid approaches and standardised evaluation methods are required for advancing this important area.

8. References

Alsubait, T., Parsia, B., & Sattler, U. (2012). Automatic Generation of Analogy Questions for Student Assessment: An Ontology Based Approach. *Research in Learning Technology*, 20, 95-101

Al-Yahya, M. (2014). Ontology-Based Multiple-Choice Question Generation. *The Scientific World Journal*, 2014.

Anderson, L., Krathwohl, R., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., Raths, J., & Wittrock, M. (Eds.) (2001). Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy. New York, NY: Longman.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5), 28-37

Biggs, J. (1999). *Teaching for Quality Learning at University: What the Student Does*. (1st ed). Open University Press

Bloom, B. S., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives. Handbook 1. Cognitive Domain*. New York: Addison-Wesley

Chickering, A. W., & Gamson, Z. F. (1987). Seven Principles for Good Practice in Undergraduate Education. *AAHE bulletin*, 3, 7

Conole, G., & Warburton, B. (2005). A review of computer-assisted assessment. ALT-J, 13(1), 17-31.

Cohen, W., Ravikumar, P., & Fienberg, S. (2003, August). A Comparison of String Metrics for Matching Names and Records. In *Proceedings of the KDD-03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, Washington, DC, 7-12 August 2003. (Vol. 3, pp. 73-78)

Collins, J. (2006). Writing Multiple-Choice Questions for Continuing Medical Education Activities and Self-Assessment Modules. *Radiographics* 2006, 26, 543-551

Cubric, M. & Tosic, M. (2011) Towards automatic generation of eAssessment using semantic web technologies, *International Journal of e-Assessment*, 1:1.

Demaidi, M. N., Gaber, M. M., & Filer, N. (2017). Evaluating the Quality of the Ontology-Based Auto-Generated Questions. *Smart Learning Environments*, 4(1), 7

Draper, S. W. (2009). Catalytic Assessment: Understanding How MCQs and EVS Can Foster Deep Learning. *British Journal of Educational Technology*, 40(2), 285-293

Fischer, S. (2001). Course and Exercise Sequencing Using Metadata in Adaptive Hypermedia Learning Systems. *Journal on Educational Resources in Computing* (JERIC), 1(1es), 5

Gagne, R. M. (1965) *The conditions of learning*. Holt, Reinhart and Winston, New York

Govindasamy, T. (2001). Successful Implementation of E-Learning: Pedagogical Considerations. *The Internet and Higher Education*, 4(3), 287-299

Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, 37(2), 337-356

Holohan, E., Melia, M., McMullen, D., & Pahl, C. (2005). Adaptive E-Learning Content Generation Based on Semantic Web Technology. In *Proceedings of Workshop on Applications of Semantic Web Technologies for e-Learning*, Amsterdam, The Netherlands, pp. 29-36

Knight, C., Gašević, D. & Richards, G. (2006). An Ontology-Based Framework for Bridging Learning Design and Learning Content. *Educational Technology & Society*, 9 (1), 23-37

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice*, *41*(4), 212-218.

Littlejohn, A. (Ed.). (2003). *Reusing Online Resources: a Sustainable Approach to E-Learning*. Psychology Press

Marzano, R. J. (2001). *Designing a New Taxonomy of Educational Objectives. Experts in Assessment*. Corwin Press, Inc., A Sage Publications Company

McKenna, P. (2019). Multiple choice questions: answering correctly and knowing the answer. *Interactive Technology and Smart Education,* 16(1), 59-73.

Mitkov, R., Le, A.H., & Karamanis, N. (2006). A Computer-Aided Environment for Generating Multiple-Choice Questions. *Natural Language Engineering* 12(2): 177-194

Mitkov, R., Le, A.H., Varga, A., & Rello, L. (2009). Semantic Similarity of Distractors in Multiple-Choice Tests: Extrinsic Evaluation. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics* (pp. 49-56). Association for Computational Linguistics

Nicol, D. (2007). E-Assessment by Design: Using Multiple-Choice Tests to Good Effect. *Journal of Further and Higher Education*, 31(1), 53-64

Papasalouros, A., Kanaris, K., & Kotis, K. (2008). Automatic Generation of Multiple-Choice Questions from Domain Ontologies. In *e-Learning* (pp. 427-434)

Paxton, M. (2001). A Linguistic Perspective on Multiple Choice Questioning, *Assessment & Evaluation in Higher Education*, 25(2), 109-119

Reeve, L., & Han, H. (2005). Survey of Semantic Annotation Platforms. In *Proceedings of the 2005 ACM symposium on Applied computing* (pp. 1634-1638). ACM Reeve and Han (2005)

Sidick, J. T., Barrett, G. V., & Doverspike, D. (1994). Three-Alternative Multiple Choice Tests: An attractive option. *Personnel Psychology*, 47(4), 829-835 Sidick, Barrett and Doverspike (1994)

Thuan, N. H., Drechsler, A., & Antunes, P. (2019). Construction of design science research questions. *Communications of the Association of Information Syst*ems, *44*(1), 20.

Tosic, M. , & Cubric, M., (2009) SeMCQ – Protégé Plugin for Automatic Ontology-Driven Multiple Choice Question Tests Generation, *11th International Protégé Conference*, June 23-26, 2009, Amsterdam.

Valaski, J., Reinehr, S., & Malucelli, A. (2017). An ontology to support the classification of learning material in an organizational learning environment: An evaluation. *Interactive Technology and Smart Education*, 14(1), 67-87.

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*, 34, 40-54

Appendix A. Evaluation questionnaire

Quality (question text):

- Low - Difficult to understand
- Medium - Easy to understand, despite incorrect use of language and grammar
- High - Easy to understand, language and grammar are correct

Quality (difficulty of the question/knowledge level)

0. Not sure what this question is testing
1. This question is testing only factual knowledge
2. This question is testing understanding
3. This question is testing application of knowledge
4. This question is testing higher order knowledge levels

Validity i.e. How well the question relates to the corresponding concept(s):

0. Doesn't relate at all
1. Relates to some extent
2. Relates quite well
3. The question represents the concept fully
4. The concept is represented by the question in full.

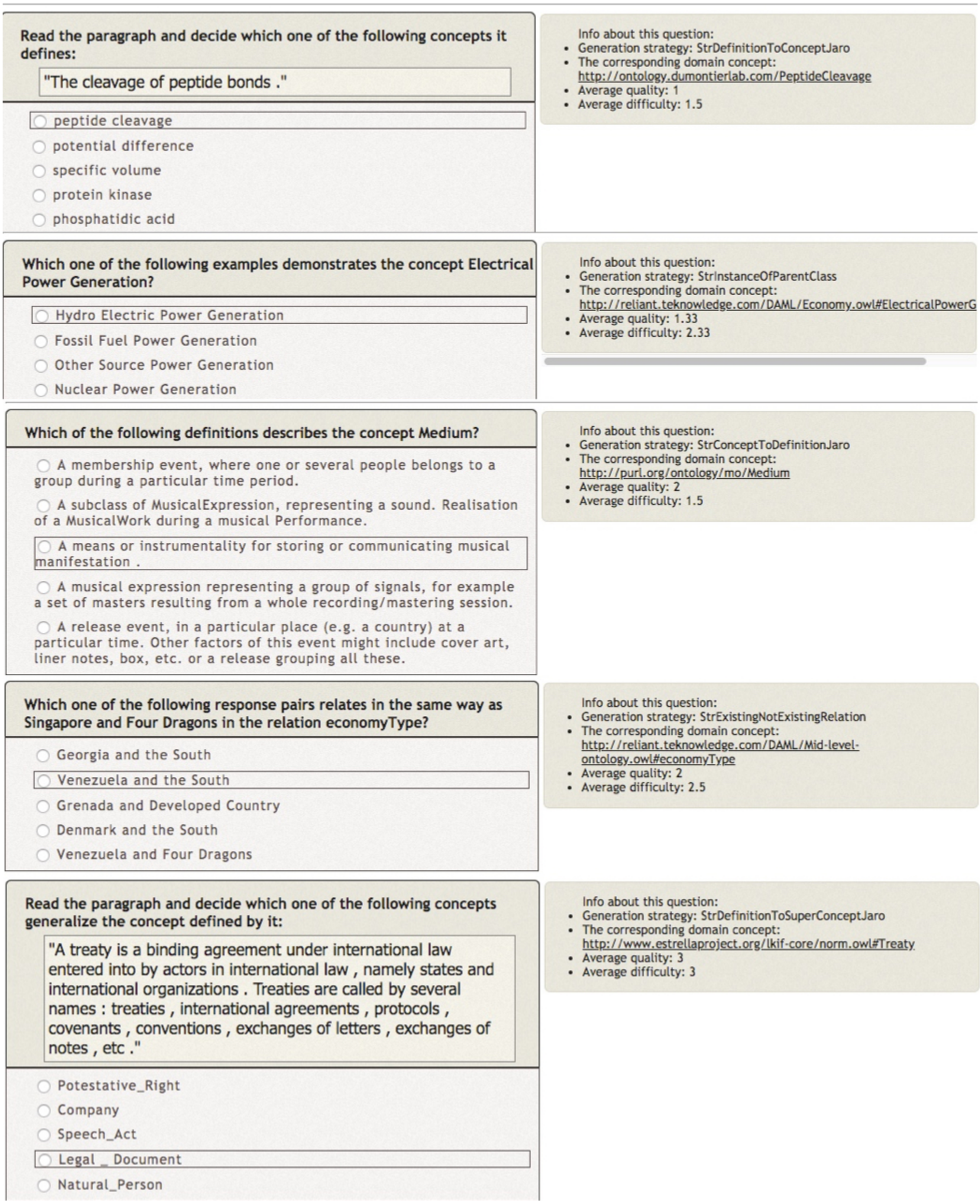
Utility:

0. This question is not usable at all
1. This question is not useable but it can serve as a base for creating another question
2. This question cannot be used due to the lack of clarity
3. This question requires language improvements to be used
4. This question can be used as is (without changes).

Any other comments:

Appendix B. Examples of evaluated questions across various evaluation ratings

Figure 5. More examples of generated questions across different evaluation ratings



Read the paragraph and decide which one of the following concepts it defines:

"The cleavage of peptide bonds ."

○ peptide cleavage
○ potential difference
○ specific volume
○ protein kinase
○ phosphatidic acid

Info about this question:

- Generation strategy: StrDefinitionToConceptJaro
- The corresponding domain concept: http://ontology.dumontierlab.com/PeptideCleavage
- Average quality: 1
- Average difficulty: 1.5

Which one of the following examples demonstrates the concept Electrical Power Generation?

○ Hydro Electric Power Generation
○ Fossil Fuel Power Generation
○ Other Source Power Generation
○ Nuclear Power Generation

Info about this question:

- Generation strategy: StrInstanceOfParentClass
- The corresponding domain concept: http://reliant.teknowledge.com/DAML/Economy.owl#ElectricalPowerG
- Average quality: 1.33
- Average difficulty: 2.33

Which of the following definitions describes the concept Medium?

○ A membership event, where one or several people belongs to a group during a particular time period.

○ A subclass of MusicalExpression, representing a sound. Realisation of a MusicalWork during a musical Performance.

○ A means or instrumentality for storing or communicating musical manifestation .

○ A musical expression representing a group of signals, for example a set of masters resulting from a whole recording/mastering session.

○ A release event, in a particular place (e.g. a country) at a particular time. Other factors of this event might include cover art, liner notes, box, etc. or a release grouping all these.

Info about this question:

- Generation strategy: StrConceptToDefinitionJaro
- The corresponding domain concept: http://purl.org/ontology/mo/Medium
- Average quality: 2
- Average difficulty: 1.5

Which one of the following response pairs relates in the same way as Singapore and Four Dragons in the relation economyType?

○ Georgia and the South
○ Venezuela and the South
○ Grenada and Developed Country
○ Denmark and the South
○ Venezuela and Four Dragons

Info about this question:

- Generation strategy: StrExistingNotExistingRelation
- The corresponding domain concept: http://reliant.teknowledge.com/DAML/Mid-level-ontology.owl#economyType
- Average quality: 2
- Average difficulty: 2.5

Read the paragraph and decide which one of the following concepts generalize the concept defined by it:

"A treaty is a binding agreement under international law entered into by actors in international law , namely states and international organizations . Treaties are called by several names : treaties , international agreements , protocols , covenants , conventions , exchanges of letters , exchanges of notes , etc ."

○ Potestative_Right
○ Company
○ Speech_Act
○ Legal _ Document
○ Natural_Person

Info about this question:

- Generation strategy: StrDefinitionToSuperConceptJaro
- The corresponding domain concept: http://www.estrellaproject.org/lkif-core/norm.owl#Treaty
- Average quality: 3
- Average difficulty: 3