Automatic Generation of Educational Quizzes from Domain Ontologies

Oscar Rodríguez Rocha, Catherine Faron Zucker

HAL Id: hal-01688798
https://hal.inria.fr/hal-01688798

Submitted on 19 Jan 2018

AUTOMATIC GENERATION OF EDUCATIONAL QUIZZES FROM DOMAIN ONTOLOGIES

O. Rodriguez Rocha[1], C. Faron Zucker[2]

[1]*INRIA Sophia Antipolis Méditerranée (FRANCE)*
[2]*University of Nice Sophia Antipolis (FRANCE)*

Abstract

Educational quizzes are very valuable resources to test or evaluate the knowledge acquired by learners and to support lifelong learning on various topics or subjects, in an informal and entertaining way. The production of quizzes is a very time-consuming task and its automation is thus a real challenge in e-Education. In this paper, we address the research question of how to automate the generation of quizzes by taking advantage of existing knowledge sources available on the Web.

We propose an approach that allows learners to take advantage of the knowledge captured in domain ontologies available on the Web and to discover or acquire a more in-depth knowledge of a specific domain by solving educational quizzes automatically generated from an ontology modelling the domain.

The implementation and experimentation of our approach is presented through the use case of a world-famous French game of manually generated multiple-choice questions.

Keywords: Semantic Web, Domain Ontologies, e-Education.

1 INTRODUCTION

Educational quizzes are very valuable resources to test or evaluate the knowledge acquired by learners and to support lifelong learning on various topics or subjects, in an informal and entertaining way. The production of quizzes is a very time-consuming task and its automation is thus a real challenge in e-Education. The research work presented in this paper contributes to answer this challenge. We address the research question of: *How can we automate the generation of quizzes by taking advantage of existing knowledge sources available on the Web?*

More specifically, we focus on the exploitation of domain ontologies available on the so-called Semantic Web. Ontologies are formal specifications of shared conceptualizations; they play a fundamental role on the Semantic Web as they enable automatic reasoning. There is a large and increasing number of domain ontologies published on the semantic Web that model various specific domains.

We propose an approach that allows learners to take advantage of the knowledge captured in domain ontologies available on the Web and to discover or acquire a more in-depth knowledge of a specific domain by solving educational quizzes automatically generated from an ontology modelling the domain.

To achieve this, our approach relies on three key features:

1. Quizzes are first formalized with Semantic Web standards: questions are represented as SPARQL queries and answers as RDF graphs. Natural language questions and answers are generated from this formalization.

2. We defined different strategies to extract multiple choice questions, correct answers and distractors from domain ontologies.

3. We defined a measure of the information content of the elements of an ontology, and of the quizzes using them.

This paper presents each of the above-cited keystones of our approach, the implementation of our approach, and its experimentation and evaluation through the use case of a world-famous French game of manually generated multiple-choice questions, which is also presented.

The remainder of this paper is structured as follows: Section 2 presents the related work and current approaches in the area automatic generation of quizzes from domain ontologies. Section 3, describes our proposed classification of the strategies for the generation of multiple choice questions through

queries, in order to align them based on their common objectives to the defined user profiles interested in learning about the given ontology. In Section 4, we define the measures of information content for the elements of a domain ontology and for the generated multiple-choice question. Our implementation and the experiments that we conducted can be found in section 5. Finally, our conclusions and future work are presented in Section 6.

2   RELATED WORK

It is possible to position our work in generating questionnaires from domain ontologies.

A set of strategies for the generation of questions, answers and distractors (false answers) from a given ontology is proposed in [1]. The strategies proposed in this work are classified into three categories: class-based, property-based and terminology-based (hierarchy of ontology). As we will explain later in this paper, our work takes these strategies as the base for the generation of the multiple-choice questions and answers.

In [2], authors present approach to automated generation of computer-assisted assessment (CAA) from Semantic Web-based domain ontologies. They propose the addition of annotations to the meta-ontology used for the generation of questions and the addition of a semantic interpretation of the mapping between the "domain" ontology and the target "question" ontology. The semantic interpretation is based on the notion of "question templates" found on the Bloom's taxonomy of educational objectives.

The OntoQue system (an engine for objective assessment item generation based on domain ontologies) is presented in [3]. The system engine uses knowledge inherent in the ontology about domain entities such as classes, properties, and individuals to generate semantically correct assessment items. Author evaluated the system with four OWL ontologies from different domains.

Authors in [4], propose an approach to generate educational assessment items using Linked Open Data (DBpedia). In their work, they describe a streamline to create variables and populate simple choice item models using the IMS-QTI standard. This work is limited to statically selecting a set of specific categories common to a set of resources from which questions, answers and distractors will be generated using SPARQL queries.

In [5] authors describe an approach for the validation of ontologies allowing to abstract the complexity of the formal languages using a set of questions and answers to which the expert is subjected. In their approach, they try to validate the conceptualization of the domain. Reasoning and verbalization techniques are used to transform the facts present in ontology into questions expressed in natural language to be evaluated by the expert. The answer to each of these questions is then processed to automatically validate or correct the ontology. Unlike this approach, ours has not yet been evaluated in a context of validation of ontologies.

An approach to the automated generation of Multiple Choice Questions (MCQs) is presented in [6]. Their approach is based on the unsupervised Relation Extraction technique used in Natural Language Processing problems. It aims to identify the most important named entities and terminology in a document and then recognize semantic relations between them, without any prior knowledge as to the semantic types of the relations or their specific linguistic realization.

In [7–9], authors present an approach to automatically generate multiple-choice questions from OWL ontologies. They describe a psychologically-plausible theory based on similarity [10] to control the difficulty of questions. They also describe a protocol to evaluate the characteristics of the generated questions such as the question difficulty.

One of the main differences with the works abovementioned, is the definition and use of information content measures of the elements of the ontology, by using these measures we can calculate the respective information content of each question, answer and distractor generated. In this way, we can automatically select a set of questions based on their information content and generate a quizz.

Our approach is based on the semantic generation of statements through SPARQL queries followed by their translation in natural language.

3 TYPOLOGY OF QUERIES FOR QUIZZ GENERATION

Our approach to automatic generate quizzes relies on the work of [1] on the generation multiple choice questions from domain ontologies through queries. It distinguishes between three different categories of queries, depending on the types of elements of the ontology that are used:

- **Class-based Strategies:** this category contains 5 strategies that generate distractors based on classes and their individuals. Distractors are formed by creating sentences in the same format as the correct answer, by choosing proper individuals or classes different from those that appear in the correct answer.

- **Property-based Strategies:** this category contains 4 strategies that create question items and distractors based on the existing relationships between individuals in the ontology. Correct answers are generated from property instances in the ontology, that is, individuals related with a given property.

- **Terminology-based Strategies:** the 2 strategies in this category are based on subsumption relationships, without dealing with ontology individuals at all.

We propose a new classification of such generation strategies based on their common objective.

This new classification facilitates the implementation and use of the strategies and more importantly, it is the basis for a future generation of personalized questions, associating each category with a specific user profile.

Table 1 shows, for each category, the type of result that is expected.

Table 1. Categories of strategies and their outputs

Category	Output
Find individuals of a given class	1 individual and 1 class
Find a data property that relates a resource and a data value	1 individual or 1 data value and 1 property
Find an object property that relates two resources	1 property and 2 individuals
Find superclasses of a given class	2 classes (1 class and its superclass)

An example of a strategy belonging to the first category and written in SPARQL is shown below.

```
SELECT  ?classA ?a (group_concat(distinct ?b ; separator = " ") AS ?d)
WHERE
{  ?classA  rdf:type      owl:Class .
   ?classB  rdf:type      owl:Class .
   ?classA  rdfs:subClassOf  ?classB .
   ?a     rdf:type      owl:NamedIndividual .
   ?b     rdf:type      owl:NamedIndividual .
   ?a     rdf:type      ?classA .
   ?b     rdf:type      ?classB
   FILTER (
     ( ( ?a != ?b ) && ( ?classB != ?classA ) ) &&
     ( ! EXISTS { ?b  rdf:type  ?classA } )
   )
}
GROUP BY ?classA ?a
ORDER BY ?classA
HAVING (?d != '')
```

4   RANKING OF QUESTIONS

We present to the user, the questions resulting from the queries made to an ontology, ranked by their *information content*. Given the wide range of questions that can be generated, their scoring and ranking is a key issue to the effectiveness of our approach.

We define the information content of a question based on the notion of information content primarily defined for concepts. According to Resnik [11], the information content (IC) of a concept represents the expressiveness. The classic measures of IC consider only the concepts that can be found in a taxonomy or in a large text corpora. Considering the evolution of ontologies and their increasing complexity, it became necessary to consider also other types of relations and elements, such as properties and property restrictions. In [12], authors consider that "the information content or expressiveness of a concept is directly proportional to the number of properties it is related to by means of property functions or property restrictions". We reuse Seddiqui's metric for the calculation of the IC of a class and we will extend it to calculate the IC of properties and instances (named individuals) as described in the following.

4.1   Information Content of Classes

Seddiqui's IC relies on the IC of concepts in a taxonomy as defined by [13]. Seco's metric relies on the hierarchical structure of a taxonomy: it states that the more hyponyms a concept has, the less information it expresses. The root of a taxonomy has no information content and its leaf nodes are those that express the maximum information. The Seco's IC is defined as follows:

$$ic_{seco}(c) = 1 - \frac{\log(hypo(c) + 1)}{\log(max_{wn})}$$

where:

$hypo(c)$

is the number of hyponyms of a given concept,

max_{wn}

is the maximum number of concepts declared in the taxonomy (in our case, ontology).

Authors in [12] consider that although many domain ontologies are not as complete as WordNet in terms of concepts and concept organization, they have good support of logical assertions to define a concept concisely. The authors capture the fact that the intrinsic source of information of a concept is its relations in terms of property functions and property restrictions, by defining the following relation-based IC metric:

$$ic_{rel}(c) = \frac{\log(rel(c) + 1)}{\log(total\_rel + 1)}$$

where:

$rel(c)$

is the number of relations of a given concept,

$total\_rel$

is the total number of relations declared in the ontology.

Seddiqui's metric of IC captures the fact that the information content of a concept depends not only on its related properties but also on its hyponyms, and therefore combines the relation-based IC with the Seco's IC:

$$ic(c) = \rho \cdot ic_{rel}(c) + (1 - \rho) \cdot ic_{seco}(c)$$

where ρ is a coefficient factor defined by the nature of the ontology. It increases when there are few concepts in the ontology and decreases when relations are inadequate to define a concept and there is a large number of concepts in the taxonomy. Formally, ρ is defined as:

$$\rho = \frac{\log(total\_rel + 1)}{\log(total\_rel) + \log(total\_concept)}$$

where:

$total\_rel$

is the maximum number of relations in the ontology,

$total\_concepts$

is the maximum number of concepts in the ontology.

We reuse Seddiqui's metric to define the information content of ontology classes, and we extend it to define the information content of ontology properties and instances (individuals).

4.2 Information Content of Properties

The information content of a property depends on its implicit relation with other classes of the ontology (of which we already know their information contents) which is captured by its domain and range. Therefore, we define the information content of a property as a function of the information content of its domain and the information content of its range:[1]

$$ic(p) = \frac{ic(domain(p)) + ic(range(p))}{2}$$

where:

$ic(domain(p))$

is the information content of the domain of the property,

$ic(range(p))$

is the information content of the range of the property.

In the particular case where no domain (resp. no range) is declared, we consider that it is *owl:Thing* (which IC is null) and the above definition still holds. In the case of a *datatype property*, we consider that the information content of a datatype is null and the above definition still holds.

4.3 Information Content of Instances

We define the information content of an instance (named individual) in an ontology as a function of the information content of its class or classes, that is, its types, multiplied by a coefficient (relations coefficient) generated by dividing the total number of relations of the given instance by the maximum number of total relations that instance in the ontology can have:

$$ic(i) = \frac{1}{n}\sum_{j=1}^{n} ic(c_{i_j}) \cdot rel_{coeff} = \frac{ic(c_{i_1}) + ic(c_{i_2}) + \cdots + ic(c_{i_n})}{n} \cdot \frac{total\_rel(i)}{max_r}$$

where $ic(c_{i_j})$ is the information content of class c_{i_j} of which i is instance of.

4.4 Information Content of Questions

We define the information content of a resulting question as the average information content of its elements: classes, properties and individuals.

$$ic(q) = \frac{1}{n}\sum_{j=1}^{n} ic(e_j) = \frac{ic(e_1) + ic(e_2) + \cdots + ic(e_n)}{n}$$

where $ic(e_j)$ is the information content of the element e_j of a question.

Questions are presented to the user ranked by their information content, in decreasing order.

---

[1] In the case where several domains (resp. ranges) are declared for a property, its resulting domain (resp. range) is the intersection of the declared classes and the IC value of its resulting domain (resp. range) is computed accordingly.

5 IMPLEMENTATION AND EXPERIMENTS

We have implemented in python the aforementioned strategies for the generation of multiple response questions and also the information content measures proposed in this paper. For both the implementations we have used the *SPARQLWrapper*[2] library to manage the communication and perform SPARQL queries to our SPARQL triple store (CORESE server[3]), *RDFLib*[4] for RDF parsing and *pandas*[5] to manipulate and perform operations to the data.

Our experimentation has been carried out within the framework of project *Intelliquizz*, which is a collaborative project between the *INRIA*[6] and *Gayatech*[7]. It aims to develop an automatic engine to generate quizzes containing multiple-choice questions, answers and distractors by initially taking as a reference the topics and the information currently present in the limited dataset of questions of *Les Incollables*, using the Web of Data as source of information.

Les Incollables[8] is a world-famous questions game in French, targeted mainly to contribute to educate children and young people by playing it. A large amount of the questions of the game are multiple-choice questions, which have been originally created manually by domain experts and written on paper cards and several non-digital formats.

As part of this project, around 160,000 questions of the game have been subjected to a digitization process. Once digitized, these questions were represented and described by means of OWL ontologies and stored in a structured data format (RDF), becoming our reference data set

For the automatic generation of quizzes based on the information contained in the reference data set, we decided to focus initially on the domain of geography. We have chosen this domain because of the large amount of related resources available on DBpedia.

As first step, it was necessary to define an ontology to which to apply the strategies and algorithms aforementioned in this paper. For this, we designed an automatic process (which will not be detailed as it is out the scope of this paper) which by means of SPARQL queries, retrieves all the questions belonging to the domain of geography present in the reference dataset. For each of these questions it obtains their natural language representation. For each representation, it performs an analysis in order to relate them to a set of DBpedia resources. Finally, from these resulting sets of resources, it extracts the classes and properties from the DBpedia ontology[9] used to represent and describe them in order to generate dynamically a smaller subset of the DBpedia ontology containing them.

We applied the quizz generation strategies described in Section 3 to the resulting subset of the DBPedia ontology, and we have applied the definitions provided in Section 4, to rank the resulting questions. For now, the questionnaires we have produced select a defined number of questions ordered based on their information content, from highest to lowest.

6 CONCLUSION AND FUTURE WORK

In this paper, we proposed an automatic process for the automatic generation of quizzes containing multiple-choice questions, answers and distractors from domain ontologies. We have also detailed how we applied our approach to generate geography questionnaires automatically, based on the information contained in the geography questions of the French game "*Les incollables*".

The most important feature of our approach is that through the proposed measures of information content of the elements of the ontology (classes, properties and instances), it is possible to measure the information content of each generated question. This makes it possible for quizzes to contain questions not only with the higher IC but also with the lower IC (to increase the difficulty of the same) or mixed.

---

[2] https://rdflib.github.io/sparqlwrapper/

[3] https://corese.inria.fr

[4] https://rdflib.readthedocs.org/

[5] http://pandas.pydata.org

[6] https://www.inria.fr

[7] http://gaya.tech/

[8] http://www.lesincollables.com

[9] http://dbpedia.org/ontology/

This opens the possibility for the generation of quizzes adapted to specific user profiles and their learning. We're currently working to provide an approach for the latter.

As another future work, with the help of experts in pedagogy from *Gayatech*, we will perform a complete evaluation of the questionnaires generated automatically for the domain of geography. Of course, future generation of questionnaires for other domains is also planned as well as an experimentation with end users that will resolve the quizzes.

REFERENCES

[1] Nunes MB, McPherson M, editors. e-Learning Automatic Generation Of Multiple Choice Questions From Domain Ontologies. 2008; IADIS; 2008.

[2] Cubric, Marija, and M Tosic. Proceedings of the 2010 International Computer Assisted Assessment Conference Towards automatic generation of e-assessment using semantic web technologies. 2010; 2010.

[3] Al-Yahya, M. IEEE 11th International Conference on Advanced Learning Technologies OntoQue: A Question Generation Engine for Educational Assesment Based on Domain Ontologies. 2011; 2011.

[4] Foulonneau M. The Semantic Web: ESWC 2011 Workshops: ESWC 2011 Workshops, Heraklion, Greece, May 29-30, 2011, Revised Selected Papers Generating Educational Assessment Items from Linked Open Data: The Case of DBpedia. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012:16–27.

[5] Ben Abacha A, Dos Reis JC, Mrabet Y, Pruski C, Da Silveira M. Towards natural language question generation for the validation of ontologies and mappings. J Biomed Semantics. 2016;7:48.

[6] Afzal N, Mitkov R. Automatic generation of multiple choice questions using dependency-based semantic relations. Soft Computing. 2014;18:1269-1281.

[7] Alsubait T, Parsia B, Sattler U. Proceedings of the 11th International Workshop on OWL: Experiences and Directions (OWLED 2014) co-located with 13th International Semantic Web Conference on (ISWC 2014), Riva del Garda, Italy, October 17-18, 2014. Generating Multiple Choice Questions From Ontologies: Lessons Learnt. 2014; 2014.

[8] Alsubait T, Parsia B, Sattler U. Knowledge Engineering and Knowledge Management: EKAW 2014 Satellite Events, VISUAL, EKM1, and ARCOE-Logic, Link\"oping, Sweden, November 24-28, 2014. Revised Selected Papers. Generating Multiple Choice Questions From Ontologies: How Far Can We Go? Cham: Springer International Publishing; 2015:66–79.

[9] Alsubait T, Parsia B, Sattler U. Ontology-Based Multiple Choice Question Generation. KI - K\"unstliche Intelligenz. 2016;30:183-188.

[10] Alsubait T, Parsia B, Sattler U. 2013 Second International Conference on E-Learning and E-Technologies in Education (ICEEE) A similarity-based theory of controlling MCQ difficulty. 2013; 2013.

[11] Resnik, P. Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1 Using Information Content to Evaluate Semantic Similarity in a Taxonomy. 1995; San Francisco, CA, USA Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.; 1995.

[12] Seddiqui, M and Masaki A. Proceedings of the Seventh Asia-Pacific Conference on Conceptual Modelling - Volume 110 Metric of Intrinsic Information Content for Measuring Semantic Similarity in an Ontology. 2010; Darlinghurst, Australia, Australia Brisbane, Australia: Australian Computer Society, Inc.; 2010.

[13] Seco N, Veale T, and Hayes J. 2004. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In 16th European Conference on Artificial Intelligence. IOS Press, 1–5.