Extraction of Relevant Resources and Questions from DBpedia to Automatically Generate Quizzes on Specific Domains

Oscar Rodríguez Rocha, Catherine Faron Zucker, Alain Giboin

Extraction of Relevant Resources and Questions from DBpedia to Automatically Generate Quizzes on Specific Domains

Oscar Rodríguez Rocha, Catherine Faron Zucker, Alain Giboin

University Côte d'Azur, CNRS, Inria, I3S, France
oscar.rodriguez-rocha@inria.fr,faron@unice.fr,alain.giboin@inria.fr

Abstract. Educational quizzes are useful not only to evaluate or test the knowledge acquired by a learner, but also to help her/him to deepen knowledge about a specific domain or topic in an informal and entertaining way. The production of quizzes is a time-consuming task that can be automated by taking advantage of existing knowledge bases available on the Web of Linked Open Data (LOD). However, automatically extracting from the LOD a knowledge graph composed by the information of a set of resources which are relevant to a given specific domain or topic, is a crucial phase for the automatic generation of quizzes.
To address this issue, we propose a heuristic that extracts from DBpedia a set of resources related to a given specific domain. Such heuristic has been implemented and used for the automatic generation of quizzes in the geography and privacy domains. We report a comparative user evaluation of it.

1 Introduction

Educational quizzes are useful not only to evaluate or test the knowledge acquired by a learner, but also to help her/him to deepen knowledge about a specific domain or topic in an informal and entertaining way. The so-called Semantic Web introduces semantics into the Web to extend its capabilities. It relies on the publication of structured data which can be viewed as a global giant knowledge graph and on ontologies which capture the relations among concepts [1] and provide the semantics that machines can understand and process. The enormous and continuous growth of this global knowledge graph makes it a rich source of structured data. As discussed in [2], the generation of quizzes *is a time-consuming task that can be automated by taking advantage of existing knowledge bases available on the Semantic Web*, for this, authors proposed an approach that relies on the work of [3] on the generation of multiple choice questions from domain ontologies through queries. However, some of those available knowledge bases (like DBpedia) may contain resources from different domains or topics, thus the automatic extraction of a knowledge graph which contains resources relevant to a specific domain or topic is a crucial phase for the automatic generation of quizzes. By *relevant resource*, we mean a resource whose

information or content is related to a specific domain and therefore it is likely to be used for the generation of questions of such domain.

The research work presented in this paper addresses the research question: *How to select a set of resources relevant to a topic or domain from a knowledge base, and extract a knowledge graph in order to be able to automatically generate quizzes from it?*

For this, we have focused our study on the DBpedia knowledge base and we have considered the definition of a topic or a domain as an input set of keywords in natural language. As a result, we propose a heuristic that selects a set of DBpedia resources that are relevant to the specified domain to generate a specific knowledge graph with their structured information, from which the questions of a quiz are generated. This heuristic finds an initial set of relevant resources through a process of entity linking and then enriches them with additional resources that are obtained through a filtering process applied to their *wikilinks*. We have carried out a comparative evaluation of this heuristic against the baseline (a heuristic that applies an entity linking process to the keywords that define a domain) in terms of relevance. In addition, we we have evaluated the relevance of the questions generated from the knowledge graphs extracted by both heuristics. By *relevant question*, we mean a question that requests information about a specific domain or allows to verify knowledge about it.

The remainder of this paper is structured as follows: In section 2, we present and detail our proposed heuristic. In section 3, we describe the implementation and evaluation of the proposed heuristic. In section 4, we present the related works. Finally, conclusions and future work are presented in Section 5.

2 Proposed approach

The selection of resources that are relevant to a specific domain or topic from a dataset is the basis for automatically generating useful quizzes for the learners. Since the knowledge bases on the Semantic Web use different ontologies and ways of structuring their data, we decided to focus our study on DBpedia since it is widely used and provides a large amount of resources from different domains.

We have considered that the simplest way to specify the domain or topic for which we want to generate quizzes, is to start with a set of keywords. From this set of keywords, in order to extract a subgraph from DBpedia with resources relevant to the described domain, we have designed our proposed heuristic, whose objective is (1), to be able to discover more relevant resources from DBpedia than with the baseline, and (2) to limit the number of non-relevant resources that may be discovered. This heuristic is inspired by the work described in [4]. We consider as a baseline, an entity linking process applied to the domain-specific keywords, such as applying DBpedia Spotlight[1] to extract a set of resources and their RDF triples to create a knowledge graph. Theoretically, it selects resources with a greater precision and lesser recall.

[1] http://www.dbpedia-spotlight.org

We will report on two different experiences: In the first one, the domain is initially specified through a set of 107 keywords resulting from a process of manual annotation of a representative set of 126 questions, extracted from the famous French game *"Les Incollables"*[2], about geography. In the second experience, the domain is initially specified through a set of 240 keywords extracted manually by an educational engineer, from resources of a MOOC about privacy[3].

2.1 Named Entity and Filtered Category Extraction

Our proposed heuristic extracts a set of DBpedia named entities R from the set of keywords that describe the targeted domain and enrich them with the *wikilinks* of the those extracted resources having relevant categories. A set of DBpedia categories C, is built with the result of a dedicated SPARQL query on DBpedia searching for the value of the *dcterms:subject* property or *dcterms:subject/skos:broader* property path of each resource in R. It considers the most relevant categories of the domain C_{topK} to filter the *wikilinks*.

We define C_{topK} as a subset of C that has k elements, which are the most relevant categories having the more related resources. The value of the number of categories k is determined by a manual analysis of the relevance of the categories with respect to the domain, allowing to discard some categories that may not be relevant to the targeted domain. For the domains of geography and privacy, the value of the threshold k was empirically fixed to 11 and 10 respectively.

As for the above-described heuristics, the set of triples describing the resources in R are stored in a named graph NG. Theoretically, this heuristic selects resources with a greater recall and lesser precision.

3 Empirical Validation of the Proposed Approach

We have applied the above described heuristics to the sets of keywords that define the two domains considered: geography and privacy.

To validate our proposed approach to extract a knowledge graph from DBpedia relevant to a given a specific domain to generate quizzes, we first evaluated the relevance of the resources selected by the baseline and our heuristic, then we measured the relevance of the questions generated from the different generated knowledge graphs.

3.1 Evaluation of the relevance of the selected resources

For the domain of *Privacy*, we have asked three students registered to the MOOC about privacy (who were also familiar with DBpedia), to evaluate the relevance of the resources selected. The list of all the resources selected by the baseline and our heuristic (and merged to avoid duplicates) was presented to them into

[2] http://www.lesincollables.com
[3] https://www.fun-mooc.fr/courses/course-v1:inria+41015+session01/about

a spreadsheet. Each resource was evaluated by the students on a scale of 1 (not at all relevant) to 5 (very relevant). Again, considering that a relevant resource is a resource related to the specific domain and therefore likely to be used in the generation of a question.

Once the relevance of the resources was evaluated, we calculated the precision and the recall of the baseline and our heuristic per user. We defined them as the proportion of relevant resources among all the resources generated by a given heuristic and the proportion of relevant resources generated by a given heuristic among all the relevant resources generated by any of the two heuristics, respectively.

According to the previous defined scale of relevancy, we considered that a resource is sufficiently relevant if its score is greater than or equal to 3.

Finally the average precision and recall (considering the three evaluators) are reported in table 1. For the domain of *Geography*, we have asked three school

	Privacy		**Geography**	
	BL	**H**	**BL**	**H**
Precision	0,567668401	0,572322652	0,957446809	0,906040268
Recall	0,34855581	0,712087542	0,25862069	0,775862069

Table 1. Precision and recall of each heuristics by domain

teachers to evaluate the relevance of the resources obtained. Similarly to the previous experimentation, a list of resources was presented to the evaluators in a spreadsheet, to be evaluated on a scale of 1 to 5.

The average precision and recall (considering also the three evaluators) are also reported in table 1.

3.2 Evaluation of the relevance of the generated questions

After having conducted the evaluation of relevance of the resource selected by each proposed heuristic, we have applied the quiz generation techniques proposed in [2], to the DBpedia subgraphs generated by the baseline and our proposed heuristic.

Finally, we have asked the evaluators to evaluate the relevance of the questions according to their corresponding domain, generated from each heuristic (on a scale of one to 5, where 5 is the most relevant). For this, they have been been provided with a list of 100 questions extracted randomly from the subgraph created for each ontology and each domain.

The results of this evaluation are shown below in table 2.

3.3 Discussion of the results

The results of the evaluation of the relevance of resources for the *Geography* domain are similar to those of the *Privacy* domain: our proposed heuristic obtains

	SubGraph BL	SubGraph H
Geography	3.59	3.5
Privacy	4.2	3.62

Table 2. Average relevance of the questions per domain and subgraph

the highest recall while keeping an accuracy not so inferior with respect to that of the baseline. Thus we can expect that our proposed heuristic is more adequate than the baseline to generate the knowledge graph from which to generate questions.

This was confirmed by analyzing the relevance of the questions generated from the two knowledge graph: the questions generated from the knowledge graph generated by the baseline were those with the highest relevance. Nevertheless, the difference in relevance with respect to the questions generated from the knowledge graph generated by our heuristic is not so great, thus our heuristic can be considered as an excellent option since it is able to discover a larger amount of related resources and therefore of questions with greater novelty.

4 Related Work

In [5], the authors propose *an approach to identify a minimal domain-specific subgraph by utilizing statistic and semantic-based metrics*. This approach targets DBpedia as a knowledge base and focuses on identifying entities and relationships strongly associated with the domain of interest. They describe the domain of interest through a main entity that represents it. In contrast, we present an approach to describe a domain in a more complete and specific way.

In [4][6][7], the authors describe an approach to exploit semantic relations stored in the DBpedia dataset to extract and rank resources related to the user context given by the keywords she enters in a search engine to formulate her query. Compared to this related work, our approach seeks to rely only on DBpedia and does not consider external sources of unstructured knowledge. Additionally, it does not currently consider estimating the *strength of the connection* between two resources linked through a *wikilink* property.

Authors in [8], present an approach to generate questions from the LOD for the History domain. In contrast to this work, our approach allows to create knowledge graphs about domains that can be defined in a more granular and specific way through keywords, and we do not rely on the use of DBpedia classes, but on entity linking combined with a process to discover related resources.

In [9], the authors present a list of state of the art works about resource discovery and graph exploration.

5 Conclusions and Future Work

In this paper, we have presented an approach for the extraction of relevant knowledge graphs from DBpedia in order to automatically generate quizzes on specific

domains. For this we have proposed and detailed a heuristic that we have further evaluated with two real life domains and educational contexts: quizzes to enrich a MOOC on privacy and quizzes to populate a serious game on geography. We have applied techniques of automatic generation of quizzes from the resulting knowledge graphs to understand the impact of the heuristics chosen to generate the input knowledge graph on the generated quizzes. The baseline heuristic consists in considering the graph of the descriptions of the named entities extracted in DBpedia from a set of keywords can be refined. The experiments showed that (1) Enriching the knowledge graph with semantically related DBpedia resources enables to increase the number of generated questions, and (2) Ranking the candidate related resources by their degree of relevancy to the domain enables to maintain the precision of the generation of questions.

As future work we will seek to improve our heuristic, by considering the participation of an expert user to the process of validating the automatically extracted categories and eventually the ranking of resources. This should improve the relevance of the generated quizzes to the domain or topic considered. Finally, we plan to evaluate the heuristic with other domains or topics.

References

1. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific american **284**(5) (2001) 28–37
2. Rodriguez Rocha, O., Faron Zucker, C.: Automatic generation of educational quizzes from domain ontologies. In: EDULEARN17 Proceedings. 9th International Conference on Education and New Learning Technologies, IATED (2017) 4024–4030
3. Papasalouros, A., Kanaris, K., Kotis, K.: Automatic generation of multiple choice questions from domain ontologies. In Nunes, M.B., McPherson, M., eds.: e-Learning, IADIS (2008) 427–434
4. Mirizzi, R., Ragone, A., Di Noia, T., Di Sciascio, E.: Semantic tags generation and retrieval for online advertising. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. CIKM '10, New York, NY, USA, ACM (2010) 1089–1098
5. Lalithsena, S., Kapanipathi, P., Sheth, A.: Harnessing relationships for domain-specific subgraph extraction: A recommendation use case. In: 2016 IEEE International Conference on Big Data (Big Data). (2016) 706–715
6. Mirizzi, R., Ragone, A., Di Noia, T., Di Sciascio, E.: Semantic tag cloud generation via dbpedia. In Buccafurri, F., Semeraro, G., eds.: E-Commerce and Web Technologies. (2010) 36–48
7. Mirizzi, R., Ragone, A., Di Noia, T., Di Sciascio, E.: Semantic wonder cloud: Exploratory search in dbpedia. In Daniel, F., Facca, F.M., eds.: Current Trends in Web Engineering. (2010) 138–149
8. Jouault, C., Seta, K., Hayashi, Y.: Content-dependent question generation using lod for history learning in open learning space. New Generation Computing **34**(4) (2016) 367–394
9. Figueroa, C., Vagliano, I., Rodríguez Rocha, O., Morisio, M.: A systematic literature review of linked data-based recommender systems. Concurrency and Computation: Practice and Experience (2015)