Translating Natural Language Competency Questions into SPARQLQueries: A Case Study

Leila Zemmouchi-Ghomari
UMBB (M'hammed Bouguerra Boumerdès University)
Boumerdès, Algeria
l_ghomari@umbb.dz

Abdessamed Réda Ghomari
LMCS Laboratory
E.S.I (National Superior School of Computer Science)
Algiers, Algeria
a_ghomari@esi.dz

***Abstract*—Ontology validation is an important part of measuring the quality of an ontology, and the best way to assure the accuracy of the knowledge encoded in the ontology. One of the earliest approaches toward ontology evaluation was the introduction of competency questions, i.e., natural language questions that the ontology should be able to answer. Since the ontology is a machine readable representation of knowledge, end-users should be able to query it using a formal language, such as SPARQL; however, translating natural language competency questions into SPARQL queries is not a trivial task. In the scope of this paper, we consider competency questions of HERO (Higher Education Reference Ontology) ontology we have developed. We translated these competency questions into SPARQL queries using a variation of a known approach.**

Keywords-Competency question; SPARQL query; Ontology validation; translation.

I. INTRODUCTION

Competency questions (CQs) [1] are the set of requirements or needs that the ontology should fulfill; they can be considered as a test collection, providing value during ontology analysis and validation [2].

According to Presutti et al. [3], CQs are used through the whole ontology development; the validation will be achieved by:

- Formalizing competency questions in the form of queries;
- Associating each query with the expected result;
- Running the queries against the ontology;
- And comparing actual with expected results.

So, in order to enable automatic evaluation with regard to competency questions, they need to be formalized in a query language. The query language has to be expressive enough to encode the competency questions appropriately.

We support the fact that SPARQL (Simple Protocol And RDF Query Language) [4] can represent a wide range of natural language questions, this language allows a high expressivity by representing and interrogating data as instances of concepts and relations defined in a reference ontology [5]. In addition SPARQL is the language proposed by W3C for querying RDF (Resource Description Framework) [6] data published on the Web.

Though translating natural language competency questions into SPARQL queries is not a trivial task [7][8]. To the best of our knowledge, automatic translation of competency questions into SPARQL queries, with the aim of validating an ontology, has not been tackled by researchers.

Although, in a more general perspective, there exist several approaches dedicated to web Question Answering (QA) area, which can potentially be exploited in addressing our specific issue. An overview of these approaches is presented in Section II. In Section III, we will describe our proposed approach. The translation process of HERO [9] competency questions into SPARQL queries is explained in Section IV and we will conclude our paper in Section V.

II. RELATED WORK

Several web QA approaches supported in most cases by platforms have been proposed to function as either natural language ontology editors, such as CNL editor [10] and OWLPATH [11], or natural language query systems like PANTO [12] and DEANNA [13]. Other approaches address this issue for a specific knowledge domain, such as: the medical domain in [14]. Table I summarizes the main features of each approach.

TABLE I. SOME WEB QA APPROACHES

Approach	Description
CNL editor	Formerly ontopath, it is composed of "OntoPath-Syntax", "OntoPath-Object" and "OntoPath-Semantic". After defining a set of concepts and relationships, the system returns the RDF ontology, and then natural language is expressed graphically by representing ontology elements, next the query is formed from the knowledge available in ontology and translated into RDF. CNL editor extends OntoPath in providing a context-free grammar with lexical dependence for defining grammars.
OWLPATH editor	It uses controlled language and grammar which are determined by question ontology. Defining the grammar using OWL ontology has two main advantages: the use of reasoners for consistency checking and the possible inclusion of restrictions in the properties of the question ontology. Thus, the grammar can take into account these restrictions while the sentence is being entered.

 ISBN: 978-1-61208-248-6

PANTO Portable Natural Language Interface to Ontologies	WordNet and string metrics algorithms are integrated into PANTO system to help make sense of the words in the NL queries and map them to the entities in the ontology. Then nominal phrases are extracted in the parse trees as pairs to form *QueryTriples*. Next, by using knowledge in the ontology, PANTO maps *QueryTriples* to *OntoTriples* which are represented with entities in the ontology. Finally, together with *targets* and *modifiers* extracted from the parse trees, *OntoTriples* are interpreted as SPARQL queries.
DEANNA Deep Answers for Naturally Asked Questions	This method is composed of six phases: first, semantic items are extracted from natural language questions then they are mapped to potential knowledge bases entities. The next phase generates candidate triples which are disambiguated in order to form semantic triples. On the basis of these triples a SPARQL query is generated.
Ben Abacha& Zweigenbaum approach: Translating Medical Questions into SPARQL Queries	This approach is composed of six phases, that is: Identifying the question type (e.g., WH: What, Who, Why, etc., Yes/No, Definition) then Determining the Expected Answer(s) Type(s) for WH questions next Constructing the question's affirmative and simplified form (new form). The following phase is Medical Entity Recognition and Relation Extraction based on the new form of the question and finally, SPARQL Query Construction

There is a limitation shared by all described approaches which is scalability, as the ontologies used for test purposes are relatively small.

Several approaches, such as DEANNA and PANTO, suppose that for every queried knowledge base, there exists a dictionary that maps questions' concepts to potential knowledge bases' semantic items, which is obviously tricky to carry out and to maintain, particularly for huge knowledge bases such as DPEDIA, Yago and Freebase. In addition, Vocabularies of the knowledge bases are controlled, so it is a challenge to correctly map users' words to vocabularies of the knowledge bases [12].

The lastly mentioned approach is limited to a particular set of questions: WH questions, except complex ones (why and when).

III. COMPETENCY QUESTIONS TRANSLATION APPROACH

Compared with web Question-Answering issue in general, our proposal tackles a specific issue. Actually, we target particular users, namely, knowledge/ontology engineers who are involved in an ontology building process. And in order to validate built ontology, they need to translate ontology specification in the form of natural language competency questions into SPARQL queries.

Expertise of these users leads us to make three assumptions before describing the proposed approach, i.e., our users are familiar with:

- Formal ontology languages (RDF or OWL) and web query languages used over ontologies/knowledge bases (SPARQL).
- Structure and vocabulary of the queried Ontology

The third assumption is related to extracted terms from competency questions which are similar to terms used to name ontology entities, according to NeOn methodology guidelines [15].

We investigated the related work (Section II) to find a methodological baseline in order to carry out a practical case study rather than a ready-to-use toolset, which has not yet been approved broadly by web QA community.

In our opinion, Ben Abacha & Zweigenbaum approach [14] fits to some extent to our needs. Actually, it is quite intuitive and relatively simple.

However, this approach is specific to the medical field, as explicitly mentioned in phase 4 of the approach; in addition, phase 2 shows that the approach focuses on a subset of WH questions which is not our intention.

Hence, we decided to slightly adapt it to our needs and the resultant approach can be summarized in five steps as illustrated by Figure 1:

1) *Identifying competency questions' categories according to expected answers' types* [14]: there are five sets of questions which are:
   a) Definition Questions: that begins with "*What is/are*" or "*What does mean*"
   b) Boolean or Yes/No Questions
   c) Factual Questions: the answer is a fact or a precise information
   d) List questions: the answer is a list of entities
   e) Complex Questions: that begins with "*How*" and "*Why*", in this case, obtaining a precise answer is almost improbable.
2) *Determining the expected (perfect or ideal) answer;*
3) *Extracting Entity or Entities from questions and their corresponding expected answers identified in 2;*
4) *Identifying answer entity type (class, data property, object property, annotation, axiom, instance) and entity location in the ontology;*
5) *Constructing the appropriate SPARQL query that gives the closest answer to the ideal answer:* based on question type identified in phase 1, question/answer entity extracted from phase 3 and its corresponding entity type/entity location in the ontology from phase 4 (as illustrated by input arrows pointing to "SPARQL Query Construction" in Figure 1).

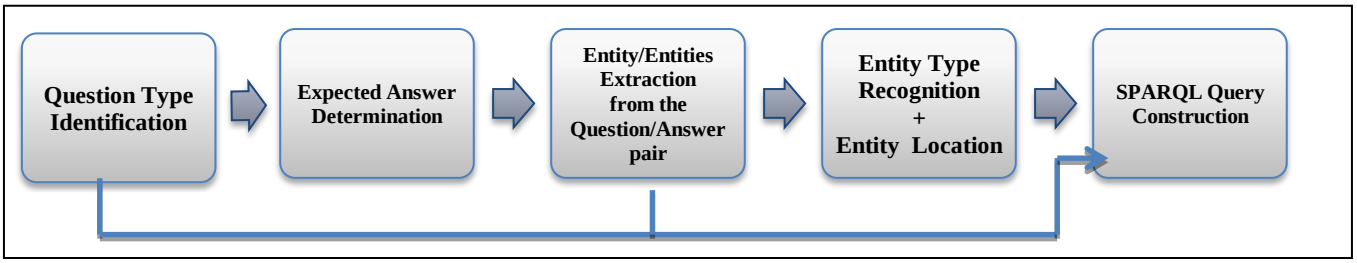

Figure 1. Competency Questions Translation Approach

IV. COMPETENCY QUESTIONS TRANSLATION PROCESS

Based on the approach description in the previous Section, we carry out the translation process of HERO ontology competency questions into SPARQL queries.

A. Identifying Question's Category

As a first attempt to detect HERO ontology requirements, we have identified eighty one (81) Competency questions (CQs) in the specification phase of HERO ontology development process; these CQs have been divided into six (6) sets according to sub-domains of higher education knowledge domain, namely: Faculty, appointments and research area, students and their life, administration, Degrees and curriculum, programs, finance, governance.

Another classification of these CQs is required according to answers types, as mentioned in the previous Section. An example of each question categoryis provided in Table II (CQs' numbering is similar to the one used in the full list of HERO CQs [16].

TABLE II. SOME EXAMPLES OF HERO CQS ACCORDING TO THEIR RESPECTIVE CATEGORIES

CQs' Categories	CQs' Examples
Definition questions	CQ59.What is a Credit?
Yes/No questions	CQ3. Must a university teacher be a researcher?
Factual questions	CQ44. What average size and duration have governing board?
List questions	CQ1. What are the possible academic ranks of a teacher?
Complex questions	CQ41.Why universities are organized into departments?

This sorting will facilitate the construction of the corresponding SPARQL queries, for example in the case of:

- *Definition question*, in our opinion, a combination of SPARQL queries can permit to provide as much information as required; we can divide this combination into five categories, to be precise:
  1) Ascendants or super classes
  2) Descendants or sub classes
  3) List of descriptive properties or data properties
  4) Relations or object properties
  5) And annotations (definition, comment, label).

The combination has not to be complete every time the definition question is met; it depends on the scope of the expected answer.

- *Yes/No questions*, in this case, ASK form of the SPARQL query will be preferred over the other forms, i.e., SELECT, CONSTRUCT and DESCRIBE, since it provides a Boolean response;
- *Factual questions versus List questions:* in the case of factual questions, we know that the query has to target one specific entity which might be a class, an instance or whatever, at the opposite of list questions where we have to obtain a number of entities as a single answer;
- *Complex questions*, often require a detailed answer, for example: in what manner things are done or causes of some phenomena. That is why we think that in most cases, corresponding ontology annotations are considered as best answers to this type of questions;

B. Determining the Expected Answer

HERO competency questions answers come from several information sources, such as: governmental academic websites, official higher education reports, experts' interviews, etc. Some examples of these answers are presented in Table III.

TABLE III. HERO COMPETENCY QUESTIONS' ANSWERS (EXCERPT)

CQs' Categories	CQs' Examples	Corresponding Answers
1	CQ59.What is a Credit?	Each course bears a specified number of credits. In general, the number of credits a course carries is determined by the number of class hours the course meets each week.
2	CQ3. Must a university teacher be a researcher?	Nearly all faculty members are expected to engage in research.

 ISBN: 978-1-61208-248-6

3	CQ44. What average size and duration have governing board?	The average size of public boards is approximately 10 people and the average size among independent (private) institutions is 30. The length of board members' terms varies from three years to as long as 12 years.
4	CQ1. What are the possible academic ranks of a teacher?	Assistant Professor, Associate Professor, Full Professor, Professor Emeritus.
5	CQ41.Why universities are organized into departments?	The basic unit of academic organization in most institutions is the department (e.g., chemistry, political science). Every department belongs to an academic field.

C. Entity Extraction from the Competency Question/Answer

From both competency questions and their expected answers, we carried out a manual extraction of relevant terms which preferably should be equivalent to some ontology entities; elsewhere the SPARQL query will not obtain an answer encoded in the ontology. This condition is valuable to warn the ontology evaluator, that it is necessary to update the ontology by adding the missing entity.

This extraction is based on a mapping between relevant terms in questions/answers pairs and their equivalent terms in ontology; it is limited to a syntactic mapping with regard to the third assumption mentioned in Section III. An excerpt of this mapping is shown in Table IV:

TABLE IV. ENTITIES' EXTRACTION FROM HERO COMPETENCY QUESTIONS AND THEIR ANSWERS (EXCERPT)

CQs' Relevant Terms	**Answers' Relevant Terms**	**Corresponding ontology terms**
CQ59…Credit?	…course … number of credits.	Course, Credit Number
CQ3. … teacher … researcher ?	engage in research	Teacher, Researcher
CQ44. …size ..duration… governing board ?	…10 …30 people …varies from three years to as long as 12 years	Size, Duration, Governing Board
CQ1. …ranks…teacher ?	Assistant Professor, Associate Professor, Full Professor, Professor Emeritus	Rank, Teacher, Assistant Professor, Associate Professor, Full Professor, Professor Emeritus
CQ41… universities … organized into departments?	… basic unit … is the department… Every department belongs to an academic field.	Higher Education Organization, Department

D. Identifying Entity Type and Entity Location

Competency questions answers must be represented in one of the allowed forms of ontology entities like: classes, data properties, object properties, axioms, instances and annotations.

SPARQL query syntax is highly dependent on the entity type of the expected answer, for example:

1) *when the answer is an INSTANCE, the SPARQL query will then be:*

*SELECT * WHERE*
*{ ?Teacher **rdf:type** HERO:Teacher . }*

2) *when the answer is a CLASS, the SPARQL query will then be:*

*SELECT * WHERE*
*{ ?subclass **rdfs:subClassOf** HERO:Student . }*

Another indispensable parameter to construct an efficient SPARQL query is the location of the expected answer in the ontology. This parameter can directly target the required information, for example: when the expected answer is located in an annotation (definition, label, comment) of a class, the SPARQL query (CQ62. What are articulation agreements?) will then be:

SELECT ?definition WHERE
*{HERO:**ArticulationAgreement rdfs:isDefinedBy** ?definition . }*

We pursue the translation process of competency questions by identifying entity types of each extracted entity from the question/answer pair and locate it in the ontology using ontology editor search function, on one hand, and on the other hand, the support of ontology engineer who knows the vocabulary and the syntax of the ontology (second assumption, Section III). The result of this identification is presented in Table V:

TABLE V. ENTITIES' TYPES AND LOCATIONS IDENTIFICATION (EXCERPT)

CQs	**Entities' Types**	**Entities' Locations in the ontology**
CQ 59	*Class*: Course *Data Property*: CourseCreditsNumber	• CourseCreditsNumber *Domain* Course
CQ 3	*Classes*: Teacher, Researcher	• Teacher *SubClassOf* Researcher
CQ 44	*Class*: Governing Board *Data Properties*: Size, Duration	• GoverningBoardSize *Domain* GoverningBoard • GoverningBoardDuration *Domain* GoverningBoard

 ISBN: 978-1-61208-248-6

CQ 1	*Class*: Teacher *Data Property*: Rank, Assistant Professor, Associate Professor, Full Professor, Professor Emeritus	• TeacherRank Domain Teacher • AssistantProfessor *SubPropertyOf* TeacherRank • AssociateProfessor *SubPropertyOf* TeacherRank • FullProfessor *SubPropertyOf* TeacherRank • ProfessorEmeritus *SubPropertyOf* TeacherRank
CQ 41	*Classes*: Higher Education Organization, Department	• Department *SubClassOf* Faculty • Faculty *SubClassOf* Role • Role *SubClassOf* HigherEducationOrganization • Department *Definition*

E. Constructing SPARQL query

Once the ideal answer identified, the equivalent entity type recognized and the localization in the ontology has been achieved; the construction of the corresponding SPARQL query can be written (facilitated by first assumption, Section III), as displayed in Table VI:

TABLE VI. SPARQL QUERIES

Competency Questions	SPARQL Queries
CQ59.What is a Credit?	SELECT ?comment WHERE { HERO:CourseCreditsNumber rdfs:comment ?comment }
CQ3. Must a university teacher be a researcher?	ASK {HERO:Teacher rdfs:subClassOf HERO:Researcher .}
CQ44. What average size and duration have governing board?	SELECT ?university ?size WHERE { ?university rdf:type HERO:HigherEducationOrganization; ?y rdfs:subClassOf ?university ; ?y HERO:GoverningBoardSize ?size }
	SELECT ?university ?duration WHERE { ?university rdf:type HERO:HigherEducationOrganization ; ?y rdfs:subClassOf ?university ; ?y HERO:GoverningBoardDuration?duration }
CQ1. What are the possible academic ranks of a teacher?	SELECT ?a ?b ?c ?d WHERE {?a rdfs:subPropertyOf HERO:TeacherRank. ?b rdfs:subPropertyOf ?a . ?c rdfs:subPropertyOf ?b . ?d rdfs:subPropertyOf ?c .}
CQ41.Why universities are organized into departments?	SELECT * WHERE { HERO:Department rdfs:subClassOf ?x ; OPTIONAL {?x rdfs:subClassOf ?y ; OPTIONAL { ?y rdfs:subClassOf HERO:HigherEducationOrganization } } }
	SELECT ?definition WHERE { HERO:Department rdfs:isDefinedBy ?definition . }

Notice that when a single SPARQL cannot provide all identified entities, it is possible to translate a competency question into several SPARQL queries, e.g., CQ41, CQ44. Another alternative could be to make an UNION between all necessary sub queries.

V. CONCLUSION AND FUTURE WORK

Natural language competency questions translation into SPARQL queries is a *sine qua non* condition for automatic evaluation of ontology requirements satisfaction.

A well defined approach of this translation process is critical for ontology evaluation area in particular and for machine readable question answering field in a more general perspective.

Based on our intuition and on some precious guidelines provided by Ben Abacha & Zweigenbaum approach [14], we achieved the translation of Higher Education Reference Ontology (HERO) competency questions into SPARQL queries.

We are conscious that our work encompasses several limitations, such as:

- Two crucial phases of our approach are entirely manual and totally dependent of user knowledge background, namely: Entity extraction from questions/answers pairs and mapping between questions/answers relevant terms and ontology entities; semi automatic support of theses phases should be considered. We suggest the use of a natural language processing tool like GATE [17] in terms extraction phase and automatic matching systems such as COMA 3.0 [18] to carry out the mapping phase.
- Weak treatment of complex questions, more precise answers are preferred to ontology annotations.

Obviously, more empirical evaluation on the approach is required to assess its performance and its effectiveness on one hand and to test HERO ontology with broader benchmark of competency questions provided by domain experts or end-users for example.

 ISBN: 978-1-61208-248-6

Despite these limitations, we are convinced that sharing experiences can significantly help research progress in web question answering processing.

VI. REFERENCES

[1] M. Gruninger and M. S. Fox, "Methodology for the design and evaluation ofontologies", IJCAI95, Workshop on Basic Ontological Issues in Knowledge Sharing. Montreal, 1995, pp. 6.1–6.10.

[2] L. Obrst, W. Ceusters, I. Mani, S. Ray, and B. Smith, The evaluation of ontologies, in Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences, C.J.O. Baker and K.-H. Cheung, Eds. New York: Springer-Verlag, 2007, pp. 139–158.

[3] V. Presutti, E. Daga, A. Gangemi, and E. Blomqvist, "Extreme design withcontent ontology design patterns", Proceedings of the workshop on ontology patterns (WOP 2009), collocated with ISWC 2009, Washington D.C, USA, Vol.516, October 2009, pp. 83-97.

[4] http://www.w3.org/TR/rdf-sparql-query/, [retrieved: Dec, 2012].

[5] L. Zemmouchi-Ghomari and A. R. Ghomari, "Reference Ontology", Fifth International Conference of Signal-Image Technology & Internet-Based Systems (SITIS), December 2009, Marrakech, Morocco, pp.485-491.

[6] http://www.w3.org/RDF/, [retrieved: Dec, 2012].

[7] D. Damljanovic, D. M. Agatonovic, and H. Cunningham, Natural language interfaces to ontologies:Combining syntactic analysis and ontology-basedlookup through the user interaction, In The Semantic Web: Research and Applications, Lecture Notes in Computer Science, Volume 6088, 2010, pp 106-120.

[8] R. J. Mooney, "Using multiple clause constructors in inductive logic programming for semantic parsing", Proceedings of the 12th European Conference on MachineLearning, Freiburg, Germany September 2001, pp. 466–477.

[9] http://sourceforge.net/projects/heronto/?source=directory, [retrieved: Dec, 2012].

[10] H. Namgoong and H. G. Kim, "Ontology-based controlled natural languageeditor using CFG with lexical dependency," Proceedings ISWC/ASWC, Vol. 4825, Lecture Notes in Computer Science, Busan, Korea, 2007, pp. 353–366.

[11] R. Valencia-García, F. García-Sánchez, D. Castellanos-Nieves, and J.T. Fernández-Breis, "OWLPath: An OWL ontology-guided query editor," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, Vol. 41, no.1, 2011, pp. 121–136.

[12] W. Chong, X. Miao, Z. Qi, and Y. Yong, "PANTO: A Portable Natural Language Interface to Ontologies", In Proceedings of the European Semantic Web Conference, volume 4519 of Lecture Notes in Computer Science, Springer-Verlag, July 2007, pp. 473-487.

[13] M. Yahya, K. Berberich, S. Elbassuoni, M. Ramanath, V. Tresp, and G. Weikum, "Deep answers for naturally asked questions on the web of data". Proceedings of the 21st international conference companion on World Wide Web, WWW '12 Companion, Lyon, France, April 2012, pp. 445-449.

[14] A. Ben Abacha and P. Zweigenbaum, "Medical Question Answering: Translating Medical Questions into SPARQL Queries", Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, Miami, Florida, USA, 2012, pp. 41-50.

[15] M. C. Suárez-Figueroa, A. Gómez-Pérez, and M. Fernández-López, The NeOn Methodology for Ontology Engineering, Book Chapter in Ontology Engineering in a Networked World, 2012, Publisher: Springer Berlin Heidelberg, pp. 9-34.

[16] http://herontology.esi.dz/content/downloads, accessed on 3 January 2012.

[17] http://gate.ac.uk/, [retrieved: Dec, 2012].

[18] http://dbs.uni-leipzig.de/Research/coma.html, [retrieved: Dec, 2012].

 ISBN: 978-1-61208-248-6