CEUR-WS.org/Vol-2288/om2018_STpaper2.pdf

Complex matching based on competency questions for alignment: a first sketch

Elodie Thiéblin, Ollivier Haemmerlé, Cassia Trojahn

IRIT & Université de Toulouse 2 Jean Jaurès, Toulouse, France
{firstname.lastname}@irit.fr

1 Introduction

A complex alignment between a source ontology o_1 and a target ontology o_2 is a set of correspondences with at least a complex correspondence. Complex correspondences (e.g., o_1*:GenusRank* $\equiv \exists$ o_2*:hasRank*.{o_2*:genus*}) involve logical constructors (e.g., property restriction) or transformation functions of literal values (e.g., string concatenation). Complex matching approaches have emerged in the literature in the last years [10, 8, 13, 6]. While some rely on statistical methods [8, 13], others rely on linguistic matching conditions [10] or knowledge rules [6]. Many of them are based on correspondence patterns [10, 8, 13]. Following a different approach, this paper proposes a complex matching approach which relies on the notion of *Competency Question for Alignment* (CQA). CQAs express the knowledge that an alignment should cover. As for ontology authoring, they take the form of NLP questions or SPARQL queries. Our approach takes as input a set of CQAs translated into SPARQL queries over the source ontology. The answer to each query is a set of instances retrieved from a knowledge base described by the source ontology. These instances are matched with those of a knowledge base described by the target ontology. The generation of the correspondence is performed by matching the graph-pattern from the source query to the lexically similar surroundings of the target instances. For example, given the source query `SELECT ?x WHERE` {`?x a` o_1`:GenusRank.`}, and an output correspondence o_1*:GenusRank* $\equiv \exists$ o_2*:hasRank*.{o_2*:genus*}, one could translate the source query into `SELECT ?x WHERE` {`?x` o_2`:hasRank` o_2`:genus.`}. Our approach was evaluated on a set of four knowledge bases about plant taxonomy.

2 Competency questions for alignment

In ontology matching system design, a question that rises is "Are there any specifications to the matching process ? If so, what are the needs/requirements that an alignment should meet ?". Few guidelines in the literature are given to characterise an alignment and/or the matching process. One of the few examples is the NeOn methodology [4], which characterises both alignment and matching process through a set of questions: i) is matching performed under time constraints ? ii) has matching to be performed automatically ? iii) must the alignment be correct ? complete ? and iv) what type of operation (merging, query, etc.) is to be performed ? Through these questions, qualitative and applicative characteristics of an alignment and the matching process are defined. However, they do not help specifying the knowledge the alignment should cover, i.e. its scope. Here, we extend the notion of “needs” for the alignment as defined in [4] by proposing the notion of *Competency Question for Alignment* (CQA).

In order to formalise the knowledge needs of an ontology, *competency questions* (CQ) have been introduced as *ontology’s requirements in the form of questions the ontology must be able to answer* [5]. Here, a CQA expresses the knowledge that an alignment should cover in the best case (if both ontologies’ scope can answer the CQA). The first difference between CQA and CQ in ontology authoring is that the scope of the CQA is limited by the intersection of its source and target ontologies’ scopes. The second difference is that this maximal and ideal alignment’s scope is not known *a priori* (as it is the purpose of the alignment). Measuring the completeness or the competency of an alignment is, however, out of the scope of this work.

Taking into account the characteristics of CQs in the literature, we adapt them for CQAs. In [9], the authors define a set of CQ characteristics (question type, element visibility, question polarity, predicate arity, modifier, domain independent element), as well as a set of competency question patterns. Inspired from the predicate arity in [9], we introduce the notion of **question arity**, which represents the arity of the expected answers to a CQA:

- A *unary* question expects a set of instances or values, e.g., “What are the genus taxa?” *(Triticum), (Anas).*
- A *binary* question expects a set of instances or value pairs, e.g., “What is the rank of a taxon?” *(Plantae, Kingdom), (Triticum, Genus).*
- A *n-ary* question expects a tuple of size 3 or more, e.g., “In which classification is the rank of a taxon defined?” *(Triticum, Genus, Linnaeus 1753), (Plantae, Kingdom, Haeckel 1866).*

Concerning the use of CQAs, they can be used for both alignment evaluation by verifying that an alignment covers a user-defined scope, as in the OA4QA task [12], and for guiding alignment creation. Our approach falls in the latter case.

3 Proposed approach

The approach takes as input a set of CQAs translated into SPARQL queries over the source ontology. The answer to each input query is a set of instances, which are matched with those of a knowledge base described by the target ontology. The matching is performed by finding the lexically similar surroundings of the target instances. Here, CQAs are limited to unary questions, (class expressions, set of instances expected), of selection type, polarity positive and no modifier. The approach is developed in 11 steps, as depicted in Figure 1:

① Extract source DL formula e_s from SPARQL CQA (e.g., o_1*:Genus*)

② Extract lexical information from the CQA, L_s set labels of atoms from the DL formula (e.g., ”Genus”, ”genre”)

③ Extract source instances $inst_s$ (e.g., o_1*:triticum*)

④ Find equivalent or similar (same label) target instances $inst_t$ to the source instances $inst_s$ (e.g. o_1:*triticum* ∼ o_2:*wheat*)

⑤ Retrieve description of target instances: set of triples and object/subject type (e.g. ⟨*(*o_2*:wheat,* o_2*:genus) :* o_2*:hasRank,* o_2*:genus:* o_2*:Rank*⟩, ⟨*(*o_2*:emmer_wheat,* o_2*:wheat) :* o_2*:hasHigherTaxon,* o_2*:emmer_wheat:* o_2*:Taxon*⟩)

⑥ For each triple, retrieve L_t labels of entities (e.g., o_2*:hasRank* → "taxonomic rank", o_2*:genus* → "genus", o_2*:Rank* → "rank")

⑦ Compare L_s and L_t using a string comparison metric (e.g., Levenshtein distance with a threshold)

⑧ Keep the triples with the summed similarity of their labels above a threshold τ. Keep the object(/subject) type if its similarity is better than the one of the object(/subject). (e.g. sim(o_2*:genus*, L_s) > sim(o_2*:Rank*,L_s) so we only keep o_2*:genus* in the triple)

⑨ Express the triple into a DL formula (e.g., ∃ o_2*:hasRank.*{o_2*:genus*})

⑩ Aggregate the formulas into an explicit or implicit form: if two DL formulas have a common atom in their right member (target member): the atoms which differed are put together (e.g., ∃ o_2*:hasRank.*{o_2*:genus*} and ∃ o_2*:hasRank.*{o_2*:kingdom*} would give 2 formulae: ∃ o_2*:hasRank.*{o_2*:genus,* o_2*:kingdom*} and ∃ o_2*:hasRank.*$\top$)

⑪ Put e_s and e_t together in a correspondence (e.g., o_1*:GenusRank* ≡ ∃ o_2*:hasRank.*{o_2*:genus*}) and express this correspondence in EDOAL

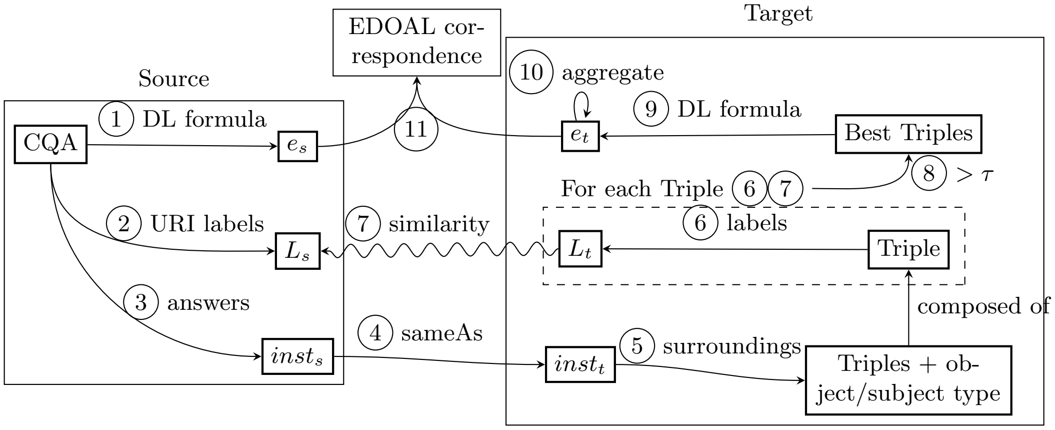

Fig. 1: Schema of the general approach.

4 Evaluation

We evaluated our approach on a set of four knowledge bases about plant taxonomy: AgronomicTaxon [11], Agrovoc [3], TaxRef-LD [7], and DBpedia [2]. All except AgronomicTaxon contain thousands of taxa ($\sim 32,000$ for Agrovoc, $\sim$

$500,000$ for TaxRef-LD, $\sim 307,000$ for DBpedia). Their instances are linked with *skos:exactMatch*, *skos:closeMatch*, *owl:sameAs* and *rdfs:seeAlso*. Two CQAs were used in the evaluation i) What are the genus taxa ? ii) What are the taxa ? Each CQA was manually translated into a SPARQL query for each ontology. All the source-target combinations of ontologies were tested, resulting in 12 alignment pairs for each CQA. For each pair, the output correspondences were manually evaluated. A correspondence was considered correct if their members are semantically equivalent. The evaluation metrics are i) precision: number of correct output correspondences / number of output correspondences and ii) top-k accuracy, as used in the evaluation of [1]: number of CQAs per pair for which at least a correct correspondence was output. As we do not compare our alignments to a reference alignment (because one would not cover all possible complex correspondences), we cannot compute recall. Table 1 presents, for each pair of ontologies and for each CQA, the number of correct correspondences out of the total number of correspondences generated by the approach. The overall precision is 32.8% (44/134) and the top-k accuracy is 83.4% (20/24). When the ontologies have a similar structure, we obtain a better precision (Agrovoc – TaxRef-LD).

	Source/Target	AgronomicTaxon	Agrovoc	TaxRef-LD	DBpedia
Genus	AgronomicTaxon		1 / 1	3 / 3	2 / 15
	Agrovoc	1 / 3		3 / 5	2 / 8
	TaxRef-LD	1 / 6	1 / 2		3 / 10
	Dbpedia	1 / 1	1 / 2	4 / 6	
Taxa	AgronomicTaxon		**0 / 4**	4 / 4	4 / 21
	Agrovoc	2 / 4		4 / 12	3 / 18
	TaxRef-LD	1 / 6	1 / 2		2 / 8
	DBpedia	**0 / 4**	**0 / 1**	**0 / 4**	

Table 1: Number of correct / number of output correspondences per CQA.

Some found correspondences were totally wrong, such as "a taxon in Agrovoc (a concept having a taxonomic rank) is something which has been represented by a statue in Wikidata" (for sake of comprehension, we express the correspondences in natural language). Other found correspondences were not precise enough such as "a taxon in Agrovoc is something having a taxon below it in a taxonomy in AgronomicTaxon", which would be correct with a subsumption relation. For some CQAs, more than one correspondence were evaluated as correct. The first reason is that some axioms of the ontology are equivalent (inverse properties, etc.). The second one is that the knowledge bases sometimes import other ontologies and instances. For example, TaxRef-LD imports data from Agrovoc, VTO and NCBI. Hence, they share common elements. Finally, as Table 1 shows, the Taxa CQA with DBpedia as source ontology does not output any correct correspondence because a taxon in DBpedia is an instance of the *dbo:Species* class. The source SPARQL query only contains this URI. Therefore, the query labels on which the lexical similarity is based are those of *dbo:Species* which do not contain anything related to *Taxon*. Most the correspondences found for this query represent the taxa having specy as taxonomic rank.

5 Conclusion and perspectives

This paper introduced the notion of competency questions for alignment (CQAs) and proposed a complex matching approach guided by CQAs. As the approach relies on the labels from the SPARQL query, the similarity of the ontologies' lexical layers impacts the output correspondences. As perspectives, we plan to perform the instance matching phase using key detection techniques, to use more linguistic evidence in the matching process, to consider binary CQAs, and work on the semantics of the confidence of complex correspondences.

Acknowledgements We would like to thank Catherine Roussey and Nathalie Hernandez for their contribution on the Agronomic dataset.

References

1. An, Y., Hu, X., Song, I.Y.: Learning to discover complex mappings from web forms to ontologies. In: ACM Conference on Information and knowledge management. pp. 1253–1262 (2012)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. The semantic web (2007)
3. Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., Keizer, J.: The agrovoc linked dataset. Semantic Web 4(3), 341–348 (2013)
4. Euzenat, J., Le Duc, C.: Methodological guidelines for matching ontologies. In: Ontology engineering in a networked world, pp. 257–278. Springer (2012)
5. Grüninger, M., Fox, M.S.: Methodology for the design and evaluation of ontologies. international joint conference on artificial inteligence. In: Workshop on Basic Ontological Issues in Knowledge Sharing. vol. 15, p. 34 (1995)
6. Jiang, S., Lowd, D., Kafle, S., Dou, D.: Ontology matching with knowledge rules. In: Transactions on Large-Scale Data-and Knowledge-Centered Systems XXVIII, pp. 75–95. Springer (2016)
7. Michel, F., Gargominy, O., Tercerie, S., Faron-Zucker, C.: A Model to Represent Nomenclatural and Taxonomic Information as Linked Data.Application to the French Taxonomic Register, TAXREF. In: S4BioDiv (2017)
8. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Discovering concept coverings in ontologies of linked data sources. In: ISWC. pp. 427–443. Springer (2012)
9. Ren, Y., Parvizi, A., Mellish, C., Pan, J.Z., van Deemter, K., Stevens, R.: Towards Competency Question-Driven Ontology Authoring. In: The Semantic Web: Trends and Challenges, vol. 8465, pp. 752–767 (2014)
10. Ritze, D., Völker, J., Meilicke, C., Šváb Zamazal, O.: Linguistic analysis for complex ontology matching. In: 5th workshop on ontology matching. pp. 1–12 (2010)
11. Roussey, C., Chanet, J.P., Cellier, V., Amarger, F.: Agronomic taxon. In: Proceedings of the 2nd International Workshop on Open Data. p. 5. ACM (2013)
12. Solimando, A., Jiménez-Ruiz, E., Pinkel, C.: Evaluating ontology alignment systems in query answering tasks. In: ISWC Posters & Demos. pp. 301–304 (2014)
13. Walshe, B., Brennan, R., O'Sullivan, D.: Bayes-recce: A bayesian model for detecting restriction class correspondences in linked open data knowledge bases. International Journal on Semantic Web and Information Systems 12(2), 25–52 (2016)