Results of CANARD in OAEI 2020*

Elodie Thiéblin[1], Ollivier Haemmerlé[2], and Cassia Trojahn[2]

[1] Logilab, France
elodie.thieblin@logilab.fr
[2] IRIT & Université de Toulouse 2 Jean Jaurès, Toulouse, France
ollivier.haemmerle@irit.fr, cassia.trojahn@irit.fr

Abstract. This paper presents the results from the CANARD system in the OAEI 2020 campaign. CANARD is a system able to generate complex alignments. It is based on the notion of competency questions for alignment, as a way of expressing user needs. The system has participated in tracks where instances are available (Populated Conference, Populated Geolink, Populated Enslaved and Taxon datasets). This is the third participation of CANARD in the OAEI campaigns.

1 Presentation of the system

The CANARD (Complex Alignment Need and A-box based Relation Discovery) system [3,4] discovers complex correspondences between populated ontologies based on Competency Questions for Alignment (CQAs). CQAs represent the knowledge needs of a user and define the scope of the alignment. They are competency questions that need to be satisfied over two or more ontologies. 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 subgraph from the source CQA to the lexically similar surroundings of the target instances.

The system source code and the configuration files are available at `https://framagit.org/IRIT_UT2J/ComplexAlignmentGenerator`.

1.1 Settings definition

Following the evaluation made in [3], the number of support instances was set to 2 instead of 10 last year to improve the runtime. The levenshtein similarity threshold was set to 0.4 like last year.

1.2 Adaptations made for the evaluation

Automatic generation of CQAs OAEI tracks do not cover CQAs i.e., the CQAs can not be given as input in the evaluation. We extended last year's query generator so that it can output binary queries. The query generator now produces three types of SPARQL queries: *Classes*, *Properties* and *Property-Value pairs*.

Classes For each *owl:Class* populated with at least one instance, a SPARQL query is created to retrieve all the instances of this class. If `<o1#class1>` is a populated class of the source ontology, the following query is created:

```
SELECT DISTINCT ?x WHERE {?x a <o1#class1>.}
```

Properties For each *owl:ObjectProperty* or *owl:Dataproperty* with at least one instantiation in the source knowledge base, a SPARQL query is created to retrieve all instantiations of this property. If `<o1#property1>` is an instantiated property of the source ontology, the following query is created:

```
SELECT DISTINCT ?x ?y WHERE {?x <o1#property1> ?y.}
```

Property-Value pairs Inspired by the approaches of [1,2,5], we create SPARQL queries of the form

- `SELECT DISTINCT ?x WHERE {?x <o1#property1> <o1#Value1>.}`
- `SELECT DISTINCT ?x WHERE {<o1#Value1> <o1#property1> ?x.}`
- `SELECT DISTINCT ?x WHERE {?x <o1#property1> "Value".}`

These property-value pairs are computed as follow: for each property (object or data property), the number of distinct object and subject values are retrieved. If the ratio of these two numbers is over a threshold (arbitrarily set to 30) and the smallest number is smaller than a threshold (arbitrarily set to 20), a query is created for each of the less than 20 values. For example, if the property `<o1#property1>` has 300 different subject values and 3 different object values (`"Value1"`, `"Value2"`, `"Value3"`), the ratio $|subject|/|object| = 300/3 > 30$ and $|object| = 3 < 20$. The 3 following queries are created as CQAs:

- `SELECT DISTINCT ?x WHERE {?x <o1#property1> "Value1".}`
- `SELECT DISTINCT ?x WHERE {?x <o1#property1> "Value2".}`
- `SELECT DISTINCT ?x WHERE {?x <o1#property1> "Value3".}`

The threshold on the smallest number ensures that the property-value pairs represent a category. The threshold on the ratio ensures that properties represent categories and not properties with few instantiations.

Implementation adaptations In the initial version of the system, Fuseki server endpoints are given as input. For the SEALS evaluation, we embedded a Fuseki server inside the matcher. The ontologies are downloaded from the SEALS repository, then uploaded in the embedded Fuseki server before the matching process can start. This downloading-uploading phase takes time, in particular when dealing with large files.

The CANARD system in the SEALS package is available at http://doi.org/10.6084/m9.figshare.7159760.v2. The generated alignments over the datasets in which CANARD performed are available at:

- **Populated Conference**: http://oaei.ontologymatching.org/2020/results/complex/popconf/results_conference.zip
- **Populated GeoLink**: http://oaei.ontologymatching.org/2020/results/complex/popgeolink/popgeolink_results_2020.zip
- **Populated Enslaved**: http://oaei.ontologymatching.org/2020/results/complex/popenslaved/popenslaved_results_2020.zip
- **Taxon**: http://oaei.ontologymatching.org/2020/results/complex/taxon/results_taxon.zip

2 Results

Please refer to http://oaei.ontologymatching.org/2020/results/complex for the results of CANARD in the OAEI 2020 campaign.

2.1 Populated Conference

Two datasets were used in the Populated Conference subtrack, one with more instances than the other. CANARD could perform all the matching tasks in the smaller dataset but timed out on 16 out of the 20 oriented pairs. This highlights one of CANARD's limitations : scalability.

For this reason, the coverage score is much lower on the large dataset than on the small one. While merging the results of all matchers by taking their best run (original, small or large dataset), CANARD obtains the best coverage score. It is the only evaluated matcher with a Coverage score higher than that of the reference simple alignment (ra1).

2.2 Populated Geolink

CANARD achieved the best relaxed-precision score (0.89) and the second best relaxed-recall score (0.54). This score however does not consider the semantics of the output correspondence. Most systems achieved a high relaxed precision score. Because of the automatic generation of CQAs, many correspondences of the form $\exists gbo{:}hasPlatformType.\{X\} \equiv gmo{:}Platform$, where X is a platform type were found.

2.3 Populated Enslaved

CANARD performed the lowest in this track out of the three evaluated complex matchers. On the enslaved-wikidata oriented pair of ontologies, CANARD found many instance links for each support answer. These links were found with literal comparison on two instances, a generic method which brings a lot of errors on a dataset with many literal information (such as dates or values). In the case of binary CQAs, as CANARD tries to find a property path between each aligned entity, the runtime exploded and had to be stopped. This shows a major flaw in CANARD that should be fixed.

2.4 Taxon

CANARD has output much more correspondences than last year. A recurring pattern was found in the correspondences: an object property from Taxref or Agrovoc is aligned to a chain of *agronomicTaxon:hasHigherRank* and *agronomicTaxon:hasLowerRank* properties. This lowered the precision score in comparison with last year's. This can be explained by:

- Wrong instance linking based on label matching regardless of the language (e.g., a plant taxon matched to a habitat)
- The computation of all possible links between the two matched instances in the target knowledge-base
- If a path is found between two matched instances, it gets a default confidence value of 0.5. If a better path is found (a path with a lexical similarity to the source property), it gets a higher value and the default path are filtered. In this case, no better path was found so all correspondences were kept.

The recall score is also lower as last year's because the system was set to use only 2 support instances this year. As the instances are not homogeneously described in each dataset, more support instances mean more chances of finding one in the target dataset which instantiate the initial knowledge need. However, CANARD still achieves the best Coverage scores.

3 General comments

CANARD relies on common instances between the ontologies. It works best with aligned instances as it will try to find lexically similar entities otherwise. Hence, when such instances are not available, the approach is not able to generated complex correspondences. Furthermore, CANARD is need-oriented and requires a set competency questions to guide the matching process. Here, these "questions" have been automatically generated based on a set of patterns.

CANARD's runtime is extremely long. It depends (among other things) on the performance of the SPARQL endpoint it interrogates and the presence (or not) of equivalent links.

However, even with generated queries (instead of user input CQAs) it obtains some of the best coverage scores.

4 Conclusions

This paper presented the adapted version of the CANARD system and its preliminary results in the OAEI 2020 campaign. This year, we have been participated in Populated Conference, Populated GeoLink, Populated Enslaved and Taxon track, in which ontologies are populated with common instances.

References

1. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Linking and building ontologies of linked data. In: ISWC. pp. 598–614. Springer (2010)
2. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Discovering concept coverings in ontologies of linked data sources. In: ISWC. pp. 427–443. Springer (2012)
3. Thiéblin, É.: Automatic Generation of Complex Ontology Alignments. (Génération automatique d'alignements complexes d'ontologies). Ph.D. thesis, Paul Sabatier University, Toulouse, France (2019), https://tel.archives-ouvertes.fr/tel-02735724
4. Thiéblin, É., Haemmerlé, O., Trojahn, C.: Generating expressive correspondences: An approach based on user knowledge needs and a-box relation discovery. In: Pan, J.Z., Tamma, V.A.M., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., Kagal, L. (eds.) The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part I. Lecture Notes in Computer Science, vol. 12506, pp. 565–583. Springer (2020), https://doi.org/10.1007/978-3-030-62419-4_32
5. 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)