Results of the Complex Taxon track of the 2018 OAEI campaign

Elodie Thiéblin and Cassia Trojahn

Institut de Recherche Informatique de Toulouse, France
elodie.thieblin@irit.fr, cassia.trojahn@irit.fr

Abstract. The Taxon dataset is part of the first complex matching track of the OAEI campaign. The particularity of this dataset is its ontologies are populated with some common instances, and that the evaluation performed over it is query oriented. This paper presents the dataset, the evaluation process and the results of the 2018 evaluation campaign.

The results presented here are the results of the Taxon dataset of the complex track of the OAEI 2018. The results have already been published on the OAEI webpage[1]. This paper gives more details about the evaluation process and the results. The Taxon track aimed at assessing the performance of matching system over large knowledge bases[2]. The evaluation was performed based on the quality of the generated alignments (in terms of precision) and on the ability of rewriting SPARQL queries using these alignments.

First, the dataset itself is presented (§1). Then, the evaluation protocol and metrics are described (§2). The system execution and results are given (§3 and §4). Finally, the results and the evaluation are discussed (§5).

1 Dataset

The dataset is composed of 4 populated ontologies whose common scope is plant taxonomy: AgronomicTaxon, AGROVOC, DBpedia, TAXREF-LD. The dataset is composed of all pairs of these four ontologies. The dataset here extends the one proposed in [2] by adding the TAXREF-LD populated ontology. A partial reference alignment used for query rewriting was manually created with the help of one expert. The correspondences only involve logical constructors, as for example:

1. $agtx{:}GenusRank \equiv agronto{:}hasTaxonomicRank.\{agv{:}c\_11125\}$
2. $agtx{:}GenusRank(x) \equiv dbo{:}genus^{-}.\top$

As these knowledge are very large, except AgronomicTaxon, as shown in Table 1, only a subset of these knowledge bases has been considered in this evaluation. This subset only considers plant taxonomy. All the plant taxa and

---

[1] http://oaei.ontologymatching.org/2018/results/complex/taxon/

[2] Comparing Taxon to LargeBio, LargeBio is richer at the TBox level, while our dataset has more instance data (except for TAXREF which has 113,252 classes.)

their information (surrounding triples, annotations, etc.) were retrieved from the SPARQL endpoint of the knowledge bases. Table 1 shows the number of plant taxa in each knowledge base. Using this first version of the dataset (some ontology files were over 400Mo), the systems loading the full ontology for the matching task encountered out of memory errors. We decided to trim the plant taxa in the knowledge base by focusing only on the *Poaceae* (*Gramineae*) family. The light version of the ontologies contains fewer instances. Even though the scope of the ontologies was reduced to plant (then *Poaceae*) taxonomy, some side-information such as the plant habitat or in which kind of beverage the plant is used is kept in the dataset.

Version	AgronomicTaxon	AGROVOC	DBpedia	TAXREF-LD
First Version	32	4,563	31,658	47,058
Light Version	32	418	4,145	3,529

Table 1: Number of plant taxa in each knowledge base of the track, in its first or light version.

2 Evaluation protocol

The main focus of the evaluation here is to be task-oriented. First, we manually evaluate the quality of the generated alignments, in terms of precision. Second, we evaluate the generated correspondences using a SPARQL query rewriting system and manually measure their ability of answering a set of queries over each dataset.

The evaluation process is done by the following steps:
- Run all the systems under the same environment
- Manually assess the output alignment to measure a precision value
- Count the number of queries a having a satisfying rewriting based on the output alignment (manual verification)

Precision We excluded from the evaluation the generated correspondences implying identical IRIs such as *owl:Class* $\equiv$ *owl:Class*. As instance matching is not the focus of this evaluation, we do not consider instance correspondences. In order to measure the precision of the output alignments, each correspondence (except if in the categories mentioned above) was manually verified and classified as true positive or false positive. A true positive correspondence has semantically equivalent members. All other correspondences are false positives. The confidence of the correspondence is not taken into account. All the systems output only equivalence correspondences so the subsumption relation was not considered.

Queries well rewritten (QWR) The queries written for the source ontology were rewritten: automatically when dealing with (1:1) or (1:n) correspondences using the system described in [3], and manually when dealing with (m:n) correspondences. The alignment used for the rewriting process is the union of the source-target and target-source alignment to take into account oriented generated alignments. If a source member included in the query is the left member of many correspondences, a target query is created for each of these correspondences. A source query is considered successfully rewritten if at least one of the target queries is semantically equivalent.

For example, a query could be "Retrieve all the genus taxa". For AgronomicTaxon, as source ontology, the corresponding SPARQL query is `SELECT ?x WHERE {?x a agtx:GenusRank.}` and the correspondences output by the systems with AGROVOC as target ontology, should be able to translate the query into: `SELECT ?x WHERE {?x agronto:hasTaxonomicRank agv:c_11125.}`

We selected 5 queries on the common scope of the ontologies: 3 of them are binary (they expect a set of value pairs), 2 are unary (they expect a set of single values). These 5 queries were chosen given the structure of the ontologies: the taxonomic rank of a taxon is represented in a certain way in an ontology and asking for *All the genus taxa* or *All the kingdom taxa* gives two very similar queries. Instead of having a query for each taxon rank, we chose only one rank. These queries were written in SPARQL for each ontology.

3 System execution

Three systems registered to the Complex track: AMLC, CANARD and XMap. AMLC needed an input simple alignment so could not be run on the Taxon track. XMap encountered an error in the execution and withdrew its application. Therefore, we chose to run all the systems which applied for the Complex, Conference and Anatomy tracks: ALOD2Vec, AML, AMLC, CANARD, DOME, FCAMapX, Holontology, KEPLER, Lily, LogMap, LogMapBio, LogMapLt, POMAP++, XMap. Out of the 14 tested systems, 7 did not output any alignment or crashed:

- ALOD2Vec encountered an exception at runtime (file lock exception);
- AMLC needed a simple alignment input;
- DOME generated empty alignments;
- FCMapX encountered an exception at runtime;
- KEPLER encountered an exception at runtime (label parsing exception);
- Lily generated empty alignments;
- XMap encountered an exception at runtime;

All the other systems (AML, CANARD, LogMap, LogMapBio, LogMapLt, POMAP++) output at least an alignment. AML only generated one alignment (AGROVOC-TAXREF) as it encountered errors during the loading phase of AgronomicTaxon and DBpedia. The only system able to output complex correspondences was CANARD.

Figure 1 shows the run-time performance of each of the systems which output at least an alignment. This performance was evaluated on a single-run on a Ubuntu 16.04 with 16GB of RAM running under a i7-4790K CPU 4.00GHz x 8 processors. CANARD was two or three times slower than the other systems.
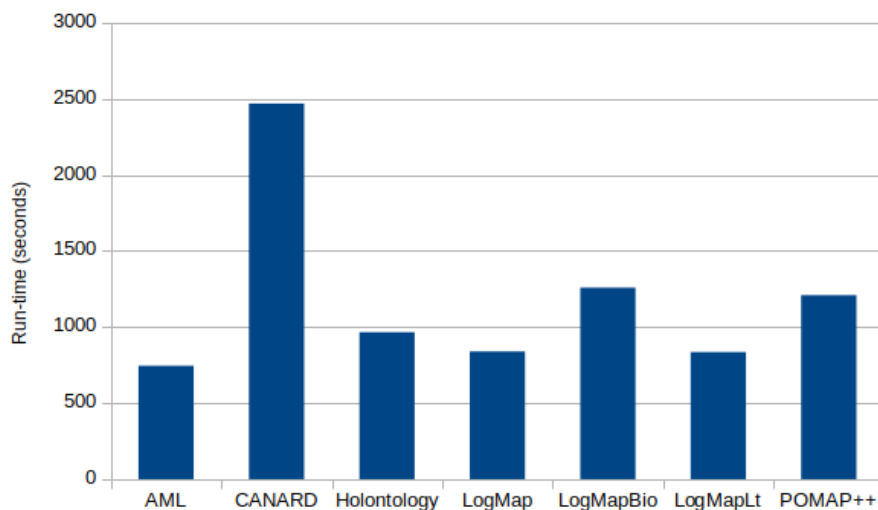

Fig. 1: Run-time of the evaluated systems

4 Results

Table 2 presents the results of the systems on the Taxon track. The table presents the run-time, the number of output correspondences, evaluated correspondence (number of output correspondences minus same IRI or instance correspondences), the number of correct correspondences. The global precision is a macro average of the precision (number of correct correspondences on all pairs divided by the number of evaluated correspondences on all pairs). The average precision is the micro average of the precision for each pair of ontologies. The number of (1:1), (1:n) and (m:n) output correspondences is also reported. Finally, the *qwr* is the percentage of queries well rewritten using the output alignments.

CANARD was the only system able to output complex correspondences on this track. LogMapLt generated the highest number of correspondences, but most were incorrect (global precision of 0.01). Overall, the best global precision is obtained by LogMap, followed by LogMapBio and POMAP++. In terms of average precision, LogMap is followed by LogMapBio and Hontology. However, these systems are not able to generate any complex correspondences.

System	Run-time (s)	output corres.	eval. corres.	correct corres.	glob. prec.	avg prec.	(1:1)	(1:n)	(m:n)	qwr
AML	745	4	1	0	0.00	0.00	4	0	0	0
CANARD	2468	142	142	27	0.19	0.20	4	66	72	**0.13**
Holontology	965	44	13	3	0.23	0.22	44	0	0	0
LogMap	839	48	19	9	**0.53**	**0.54**	48	0	0	0.07
LogMapBio	1258	45	17	5	0.29	0.28	45	0	0	0
LogMapLt	834	5064	1920	10	0.01	0.16	5064	0	0	0.10
POMAP++	1208	33	8	2	0.25	0.14	33	0	0	0

Table 2: Results. *corres.* stands for correspondence, *eval.* for evaluated, *qwr* for queries well rewritten, *glob. prec.* for global precision, *avg prec.* for average precision

Looking for the query rewriting task, only 3 systems generated alignments which could be used to rewrite queries: CANARD, LogMap and LogMapLt. CANARD's alignment was used to rewrite the most queries (best *qwr*).

Out of the 3 binary queries $\times$ 12 pairs of ontologies $=$ 36 binary query cases, only 8 could be rewritten using simple alignments including 4 with simple equivalence alignments. These 4 query cases could be successfully rewritten by LogMapLt's output, and 2 from LogMap's. CANARD did not output any alignment for the binary cases.

Out of the 2 unary queries $\times$ 12 pairs $=$ 24 unary query cases, only 8 could be dealt with simple correspondences. 2 of these 8 cases involved an instance correspondence (which was out of the scope of the evaluation). LogMap and LogMapLt could both deal with these 2 cases. CANARD could deal with 6 unary cases needing a complex correspondence and 2 needing simple correspondences.

Overall, the query cases needing simple alignments were rather well covered by simple matchers: 42% (4+2/6+8) of them for LogMapLt, 60% (4+2/4+6) if we consider only simple equivalence and leave out the queries needing instance matching; 29% and 40% for LogMap. For the query cases needing complex correspondences, (0+6/28+16) 14% were covered by CANARD. For all the query cases, the CANARD system could provide an answer to 8 query cases out of the 36 + 24 = 60 cases.

Figure 2 shows that all systems (except CANARD) output over 60% of same IRIs or instance correspondences.

5 Discussion

With respect to the generated alignments, we could observe that the CANARD system generates, for some pairs, more than one correct correspondence, as DBpedia has various ways of expressing the knowledge. For example, *agtx:GenusRank* $\equiv \exists$ *dbo:genus^{-}.$\top$* or *agtx:GenusRank* $\equiv \exists$ *dbp:genus.$\top$*. It shows the ability of
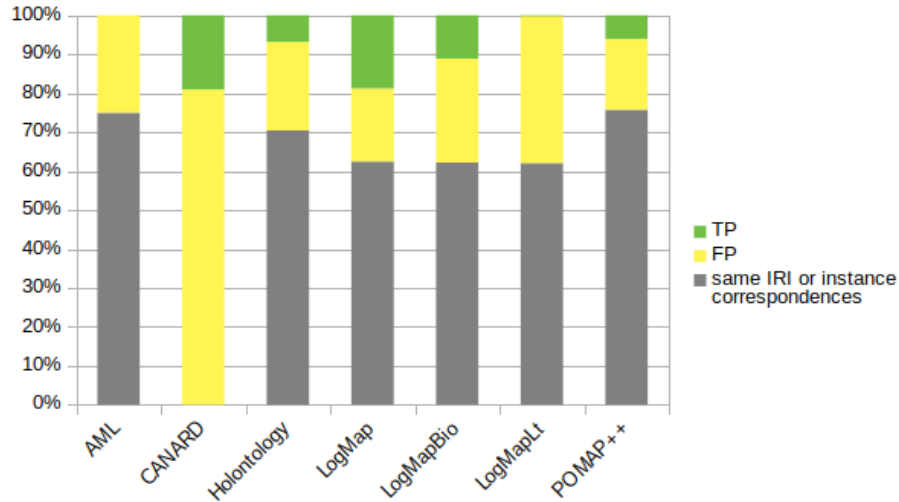
Fig. 2: Proportion of correspondences in the evaluation: not considered, incorrect and correct

the system to detect complex correspondences that express the different ways the knowledge can be expressed across datasets.

We could also observe that LogMapLt outputs a high number of heterogeneous correspondences such as class to instance, object property to class, etc. These correspondences were considered in the evaluation. Some of these correspondences were interesting such as *dbo:genus* ≡ *txrfrk:Genus* (*dbo:genus* is an object property and *txrfrk:Genus* is an instance) but as they do not provide the right semantics, they were classified as false positive. We note that for the pair DBpedia-TAXREF, 935 correspondences were found between instances and the class *dbo:Species* and 935 correspondences were found between instances and the object property *dbo:species*.

Looking to the ability of rewriting queries using the generated alignments, simple matchers could cover the most cases involving such simple correspondences. However, the main take-away of this first evaluation is that simple alignments are sometimes not sufficient for some applications. Only the CANARD system could output complex correspondences but limited to the unary queries.

With respect to the technical environment, the SEALS system as it is now was probably not adapted to deal with big knowledge bases as the loading phase got very slow. The use of SPARQL endpoints instead of two ontology files would make more sense given that many LOD repositories provide one. As the evaluation was based on a set of queries, the systems could receive the queries as input in order to limit the search space and therefore gain in efficiency. The correspondences output by the systems were expressed in the Alignment format for the simple alignments and in EDOAL [1] for the complex alignments. However, the

EDOAL language is a little limited to represent heterogeneous correspondences (correspondences which can not be represented in DL). For example, the taxonomic rank property of a taxon can be represented with an instance linked by a property, a class or a property pointing to this taxon. SPARQL could be a more expressive alignment format in future works.

References

1. Euzenat, J., Scharffe, F., Zimmermann, A.: Expressive alignment language and implementation (2007)
2. Thiéblin, E., Amarger, F., Hernandez, N., Roussey, C., Trojahn, C.: Cross-querying lod datasets using complex alignments: An application to agronomic taxa. In: Research Conference on Metadata and Semantics Research. pp. 25–37. Springer (2017)
3. Thiéblin, E., Amarger, F., Haemmerlé, O., Hernandez, N., Trojahn, C.: Rewriting SELECT SPARQL queries from 1:n complex correspondences. In: 11th ISWC workshop on ontology matching. pp. 49–60 (2016)