EXONA Results for OAEI 2015

Syrine Damak, Hazem Souid, Marouen Kachroudi
and Sami Zghal

Université de Tunis El Manar, Faculté des Sciences de Tunis, LIPAH Programmation
Algorithmique et Heuristique, 2092, Tunis, Tunisie;
damaksyrine@gmail.com
swdhazem@gmail.com
marouen.kachroudi@fst.rnu.tn
sami_zghal@planet.tn

Abstract. This paper presents the results of EXONA in the Ontology Alignment Evaluation Initiative (OAEI) 2015. EXONA is an automatic instance-based ontology alignment systems in which we parse ontology as first step. In the second step, we index instances of the first ontology. These indexed instances will be applied for the querying phase. In the last step, our system aligns instances based by aggregating score of different terminological matchers. We first describe the overall framework of our matching System (EXONA) then we detail the techniques used in the framework for instance matching. Last, we give a thorough analysis on our results and discuss some future work on our system. It's our first participation in the OAEI instance matching, the results are good in terms of recall, precision and F-measure.

1 Presentation of the system

Ontology matching is a key interoperability enabler for the semantic web, as well as a useful tactic in data integration tasks. Knowledge about one object may be contained in multiple and different knowledge bases. Therefore, a lot of work has already been built to obtain more complete knowledge about things existing in different domains. This is in order to exceed the area of divergence obstacle, by creating cross-domain knowledge.

Accordingly, it's strongly recommended to focus on the more active element of ontology which it called instance. Many instances matching approaches have been proposed, and among which is ours. In fact, our system is proposed for large scale instance matching. It operates on three successive modules, namely : transformation, indexation and correspondence. Transformation consists in transforming separately both of knowledge bases on an exploitable form and then creating our own instance object as a profession object. Indexation is the process of indexing instances of knowledge base; only instances of the source base knowledge have to be indexed. Correspondence consists in querying the index already built. This request contains instances of the target knowledge base non indexed.

In order to solve the problem of large knowledge bases, we propose an index by concept to minimise the area and the time expended on searching instances behind the request technique.

1.1 State, purpose, general statement

This section describes the overall framework of EXONA . Our system includes three modules, i.e., *transformation, indexation and correspondence*.

The system proposed operates in three successive modules, each of those is branched into two phases. The system begins with transforming the knowledge bases into two independent graphs, those graphs formed by OWL nodes. After having constructed these graphs, it's time instance objects be created. The construction of instance object appealed the neighbourhood technical in which neighbourhood spread by similarity is done. This technique aims to enrich instance object by neighbouring instances with which a high similarity exists. We proceed after that by a terminological normalization of instances. This normalisation is compulsory for the indexation of instances as well as for the similarity calculation. We index after that instances of the source data knowledge. It 's not a blind indexation. In fact it is done by concept. This is in order to provide optimal search fields oriented concept. After index creation, it's time to query this index. The request emitted contains instances from the target knowledge base. Those instances have to be normalized before being passed through the request. After querying the index, a candidate set is returned. Each candidate pair is accompanied by Ã score indicating its rank behind the rest of pair. To identify the pair of instance that have to be aligned, we filter this set of candidate set by saving only the two best pairs. As a verification process, we calculate terminological similarity. This latter want to be combined by the score given on requesting phase. Those similarities are aggregated then to identify the pair of instances to be aligned having the higher similarity score.

1.2 Specific techniques used

The process of EXONA system consists in the following three successive modules, namely : transformation, indexation and correspondence.

1. **Transformation module**
   This module is branched into two phases, namely : graph construction and instance creation.
   - *Graph construction* : As input, our system receive two OWL files. Those files are transformed into two independent OWL graphs. Graphs are more adequate representation ensuring highlighting of information.
   - *Instance creation* : Instance object is an object formed by an identifier and a content. It is identify by an URI; its content is formed by a set of information which makes it an autonomous entity. It contains the list

of neighbouring instances spread by similarity. Moreover, it contains the RDF triplet of this one. To calculate similarity, we have used $Edit-distance$.

2. **Indexation module**

   This module is formed by two phases, namely : pretreatment and instance creation.
   - *pretreatment* : In this phase we remove special symbols like "Â£.*-", etc. and stop words like "the", etc. standardization of case, etc. This pretreatment serves then in the requesting phase.
   - *Indexation* : indexation aims to index instances of the source data knowledge. Each document is identfied by the URI of the instance and contains the content of instance formed only by its data property and the data property of the set of neighbouring instances spread by similarity.

3. **Correspondence module**

   This module is divided into two phases, namely : querying and Filtering and matches identification.
   - *Querying* : querying phase has as input instances of the target knowledge base. Those instances have to be pretreated before the process of research on the index. This phase generates as an output a list of candidates accompanied by their score.
   - *Filtering and matches identification* : During this phase, the system takes the two candidates having the highest score. Then, it calculates terminological similarity of this pair of candidate. EXONA system identifies the pair of instances to be aligned by aggregating similarities with the one given by the search process. The pair of instance to be aligned is the one having the best score.

1.3 Adaptations made for the evaluation

We have changed the version of Lucene from 4.10.2 to 3.6. In fact, our first one needs specific adaptations to be accepted by 2015 evaluation campaign.

1.4 Link to the set of provided alignments (in align format)

http://www.mediafire.com/download/b3vx3zio02br45y/$EXONA_OAEI_2$015.zip

2 Results

The instance matching 2015 track contains two subtasks. Each task is articulated in two tests with different scales (i.e., number of instances to match): i) Sandbox (small scale). It contains two datasets called source and target as well as the set of expected mappings (i.e., reference alignment). ii) Mainbox (medium scale).

2.1 Author Disambiguation Task

The goal of the Author Disambiguation Task is to link OWL instances referring to the same person (i.e., author) based on their publications. This task is done with the two datasets previously invoked.

Sandbox task The Sandbox test aims to evaluate behaviour of our system with Ã small scales. Table 1 below presents the results obtained by running EXONA on the instance matching track of OAEI campaign 2015 done with the Sandbox task.

	Precision	**Recall**	**F-measure**
EXONA	0.941	0.941	0.941
InsMT+	0.834	0.705	0.764
Lily	0.981	0.981	0.981
LogMap	0.994	0.906	0.948
RiMOM	0.929	0.929	0.929

Table 1. Results of Author Disambiguation Task for Sandbox task

Mainbox task This task is also done with Mainbox task. The gaol of this test is to evaluate the behaviour of our system in large scale. Table 2 below presents the results obtained by running EXONA on the instance matching track of OAEI campaign 2015 done with the Mainbox task.

	Precision	**Recall**	**F-measure**
EXONA	0.0	0.0	NaN
InsMT+	0.76	0.665	0.709
Lily	0.964	0.964	0.964
LogMap	0.996	0.831	0.906
RiMOM	0.911	0.911	0.911

Table 2. Results of Author Disambiguation Task for Mainbox task

2.2 Author Recognition Task

The goal of Author Recognition Task is to associate a person (i.e., author) with the corresponding publication report containing aggregated information about the publication activity of the person, such as number of publications, h-index, years of activity, number of citations. This task is done also with the two datasets previously invoked.

Sandbox task The Sandbox test aims to evaluate behaviour of our system with Ã small scales. Table 3 below presents the results obtained by running EXONA on the instance matching track of OAEI campaign 2015 done with the Sandbox task.

	Precision	**Recall**	**F-measure**
EXONA	0.518	0.518	0.518
InsMT+	0.556	0.059	0.106
Lily	1.0	1.0	1.0
LogMap	1.0	1.0	1.0
RiMOM	1.0	1.0	1.0

Table 3. Results Author Recognition Task for Sandbox task

Mainbox task This task is also done with Mainbox task. The goal of this test is to evaluate the behaviour of our system in large scale. Table 4 below presents the results obtained by running EXONA on the instance matching track of OAEI campaign 2015 done with the Mainbox task.

	Precision	**Recall**	**F-measure**
EXONA	0.409	0.409	0.409
InsMT+	0.246	0.028	0.05
Lily	0.999	0.998	0.999
LogMap	0.999	1.0	0.999
RiMOM	0.999	0.999	0.999

Table 4. Results Author Recognition Task for Mainbox task

3 Conclusion

Exona participation in OAEI 2015 is encouraging although its participation is restricted to a few sub-cases, as it supplies good metric values in the two considered cases. Results reflects some strengths and some good aspects that need to be improved.

References

1. Euzenat, J., Shvaiko, P.: Ontology Matching (Second Edition). Springer-Verlag, Heidelberg (DE) (2013)