LN2R – a knowledge based reference reconciliation system: OAEI 2010 Results

Fatiha Saïs[1], Nobal Niraula[2], Nathalie Pernelle[1], Marie-Christine Rousset[3]

[1] LRI (Paris-Sud 11 University & CNRS) / INRIA Saclay Parc-Club Orsay Univ.
Bat. G ,4, rue Jacques Monod F-91893 Orsay Cedex, France
{fatiha.sais, nathalie.pernelle}@lri.fr
[2] University of Memphis, TN, USA
nb.niraula@gmail.com
[3] LIG - Laboratoire dInformatique de Grenoble
BP 72, 38402 St MARTIN DHERES, France
Marie-Christine.Rousset@imag.fr

Abstract. This paper presents the first participation of LN2R system in IM@OAEI2010, the Instance Matching track of Ontology Alignment Evaluation Initiative 2010 Campaign. In particular, we participated in OWL data track by performing LN2R system on Person-Restaurant data set. We obtained very good results on person data sets and reasonable results on restaurant data set.

1 Presentation of the system

To design a semantic information integration system, we are faced to two reconciliation problems. First, the schema (or ontology) reconciliation which consists in finding mappings between elements (concepts or relations) of two schemas or two ontologies (see [1, 2] for surveys). The second problem concerns data reconciliation (named reference reconciliation) which consists in comparing data descriptions and deciding whether different descriptions refer to the same real world entity (e.g. the same person, the same article, the same gene). The problem of reference reconciliation is very critical, since it impacts data quality and data consistency [3].

In LN2R system, we address only the problem of reference reconciliation. There are several kinds of reference reconciliation approaches: knowledge-based, similarity-based, probabilistic, supervised, etc.[4]. In this paper we focus our study on reference reconciliation approaches that are *informed* and *global*. *Informed* approaches are those which exploit knowledge that is declared in the ontology to reconcile data. Reference reconciliation approaches are said *global* when they exploits the dependencies possibly existing between reference reconciliations [5, 6]. Such approaches use attribute values describing the data but also references that are related to the considered data [5, 6]. For example, the reconciliation between two scientists can entail the reconciliation between their two affiliated universities. Such dependencies result from the semantics of the domain of interest.

1.1 State, purpose, general statement

The reference reconciliation system (LN2R) that we have tested in IM@OAEI2010 campaign is knowledge-based, unsupervised and based on two methods, a logical one called L2R and a numerical one called N2R. The Logical method for Reference Reconciliation (L2R) is based on the translation in first order logic (Horn rules) of some of the schema semantics. In order to complement the partial results of L2R, we have designed a Numerical method for Reference Reconciliation (N2R). It exploits the L2R results and allows computing similarity scores for each pair of references.

Reference reconciliation problem. Let $S1$ and $S2$ be two data sources which conform to the same OWL ontology. Let $I1$ and $I2$ be the two reference sets that correspond respectively to the data of $S1$ and $S2$. The problem consists in deciding whether references are reconciled or not reconciled. Let $Reconcile$ be a binary predicate. $Reconcile(X, Y)$ means that the two references denoted by X and Y refer to the same world entity. The reference reconciliation problem considered in L2R consists in extracting from the set $I1 \times I2$ of reference pairs two subsets REC and $NREC$ such that: $REC = \{(i, i), Reconcile(i, i)\} \quad and \quad NREC = \{(i, i), \neg Reconcile(i, i)\}$

The reference reconciliation problem considered in N2R consists in, given a similarity function $Simr : I1 \times I2 \rightarrow 0..1$, and a threshold Trec (a real value in 0..1 given by an expert, fixed experimentally or learned on a labeled data sample), computing the following set:
$REC_{N2R} = \{(i, i') \in (I_1 \times I_2) \backslash (REC \cup NREC), s.t. Sim_r(i, i') > T_{rec}\}$

1.2 Specific techniques used

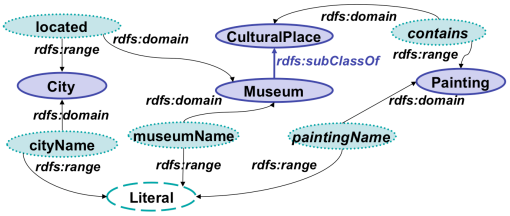
In the following, we will present some details on the knowledge-based reference reconciliation system (LN2R). First, we will show through an example the ontology and the kind of knowledge that we use. Second, we give a brief presentation of the two methods L2R and N2R of reference reconciliation.

The ontology and its constraints The considered OWL ontology consists of a set of classes (unary relations) organized in a taxonomy and a set of typed properties (binary relations). These properties can also be organized in a taxonomy of properties. Two kinds of properties can be distinguished in OWL: the so-called relations (in OWL abstractProperty), the domain and the range of which are classes and the so-called attributes (in OWL objectProperty), the domain of which is a class and the range of which is a set of basic values (e.g. Integer, Date, Literal).

We allow the declaration of constraints expressed in OWL-DL or in SWRL in order to enrich the domain ontology. The constraints that we consider are of the following types:

- Constraints of disjunction between classes: DISJOINT(C,D) is used to declare that the two classes C and D are disjoint.
- Constraints of functionality of properties: PF(P) is used to declare that the property P (relation or attribute) is a functional property.

(a) (b)



Source S1:
MuseumName(S1_m1,"Le Louvre"); Contains(S1_m1,S1_p1); Located(S1_m1,S1_c1); CityName(S1_c1,"Paris"); PaintingName(S1_p1, "La Joconde");

Source S2:
MuseumName(S2_m1,"muse du Louvre"); Located(S2_m1,S2_c1); Contains(S2_m1,S2_p1); Contains(S2_m1,S2_p2);CityName(S2_c1,"Ville de paris"); PaintingName(S2_p1, "Abricotiers en fleurs"); PaintingName(S2_p2,"Joconde");

Fig. 1. (a) an extract of cultural place ontology, (b) an extract of RDF data

– Constraints of inverse functionality of properties: PFI(P) is used to declare that the property P (relation or attribute) is an inverse functional property. These constraints can be generalized to a set $\{P_1, \ldots, P_n\}$ of relations or attributes to state a combined constraint of inverse functionality that we will denote PFI($P_1, \ldots, P_n$). For example, $PFI(located, name)$ expresses that one address and one name cannot be associated to several cultural places (i.e. both are needed to identify a cultural place).

Data description and their constraints A piece of has a reference, which has the form of a URI (e.g. `http://www.louvre.fr,NS-S1/painting243`), and a description, which is a set of RDF facts involving its reference. An RDF fact can be: either (i) a class-fact $C(i)$, where C is a class and i is a reference, (ii) a relation-fact $R(i1, i2)$, where R is a relation and $i1$ and $i2$ are references, or (iii) an attribute-fact $A(i, v)$, where A is an attribute, i a reference and v a basic value (e.g. integer, string, date).

The data description that we consider is composed of the RDF facts coming from the data sources enriched by applying the OWL entailment rules. We consider that the descriptions of data coming from different sources conform to the same OWL ontology (possibly after schema reconciliation). Figure 1 (b), provides examples of data coming from two RDF data sources S1 and S2, which conform to a same ontology describing the cultural application previously mentioned.

L2R: a Logical method for Reference Reconciliation L2R [7] is based on the inference of facts of reconciliation (Reconcile(i,j)) and of non-reconciliation ($\neg Reconcile(i', j')$) from a set of facts and a set of rules which transpose the semantics of the data sources and of the schema into logical dependencies between reference reconciliations. Facts of synonymy (SynVals(v1,v2)) and of no synonymy ($\neg SynVals(u1, u2)$) between basic values (strings, dates) are also inferred. For instance, the synonymy SynVals("JoDS", "Journal of Data Semantics") may be inferred.

The L2R distinguishing features are that it is global and logic-based: every constraint declared on the data and on the OWL ontology is automatically translated into first-order logic Horn rules (rules for short) that express dependencies between reconciliations. For instance, the following rule R translates the knowledge that the two classes $Museum$ and $City$ are disjoint, $R : Museum(X) \wedge City(Y) \Rightarrow \neg Reconcile(X, Y)$

To deduce all the (non) reconciliation and (non) synonymy facts from the knowledge base, we use a logical reasoning based on the unit-resolution inference rule. The advantage of such a logical approach is that if the data are error-free and if the declared constraints are valid, then the reconciliations and non-reconciliations that are inferred are correct, thus guaranteeing a 100 % precision of the results.

N2R: a Numerical method for Reference Reconciliation N2R [5] has two main distinguishing characteristics. First, it is fully unsupervised: it does not require any training phase from manually labeled data to set up coefficients or parameters. Second, it is based on equations that model the influence between similarities. In the equations, each variable represents the (unknown) similarity between two references while the similarities between values of attributes are constants. These constants are obtained, either (i) by exploting a dictionnary of synonyms (e.g. WordNet thesaurus, the dictionnary of synonyms generated by L2R [7]); or (ii) by using standard similarity measures on strings or on sets of strings. Furthermore, ontology and data knowledge (disjunctions and UNA) is exploited by N2R in a filtering step to reduce the number of reference pairs that are considered in the equation system. The functions modeling the influence between similarities are a combination of maximum and average functions in order to take into account the constraints of functionality and inverse functionality declared in the OWL ontology in an appropriate way.

The equations modeling the dependencies between similarities. For each pair of references, its similarity score is modeled by a variable x_i and the way it depends on other similarity scores is modeled by an equation: $x_i = f_i(X)$, where $i \in [1..n]$ and n is the number of reference pairs for which we apply N2R, and $X = (x_1, x_2, \ldots, x_n)$. Each equation $x_i = f_i(X)$ is of the form: $f_i(X) = max(f_{i-df}(X), f_{i-ndf}(X))$

The function $f_{i-df}(X)$ is the maximum of the similarity scores of the value pairs and the reference pairs of attributes and relations with which the i-th reference pair is functionally dependent. The maximum function allows propagating the similarity scores of the values and the references having a strong impact. The function $f_{i-ndf}(X)$ is defined by a weighted average of the similarity scores of the values pairs (and sets) and the reference pairs (and sets) of attributes and relations with which the i-th reference pair is not functionally dependent. See [5] for the detailed definition of $f_{i-df}(X)$ and $f_{i-ndf}(X)$.

Iterative algorithm for reference pairs similarity computation. Solving this equation system is done by an iterative method inspired from the Jacobi method [8], which is fast converging on linear equation systems. To compute the similarity scores, we have implemented an iterative resolution method. At each iteration, the method computes the variable values by using those computed in the precedent iteration. Starting from an initial vector $X^0 = (x_1^0, x_2^0, ..., x_n^0)$, the value of the vector X at the k-th iteration is

obtained by the expression: $X^k = F(X^{k-1})$. At each iteration k we compute the value of each x_i^k: $x_i^k = f_i(x_1^{k-1}, x_2^{k-1}, ...x_n^{k-1})$ until a fix-point with precision ϵ is reached. The fix-point is reached when: $\forall i, |x_i^k - x_i^{k-1}| <= \epsilon$.

The similarity computation is illustrated by the equation system (see Table 1) obtained from the data descriptions shown in the example 1.

- x_1 = Sim_r(S1_m1, S2_m1) ; Sim_v("Le louvre", "Musee du louvre") = 0.68
- x_2 = Sim_r(S1_p1, S2_p1) ; Sim_v("La Joconde", "Abricotiers en fleurs") = 0.1
- x_3 = Sim_r(S1_p1, S2_p2) ; Sim_v("La Joconde", "Joconde") = 0.9
- x_4 = Sim_r(S1_c1, S2_c1) ; Sim_v("Paris", "Ville de Paris") = 0.42

The weights are computed in function of the number of common attributes and common relations of the reference pairs. The weights used in the value computation of the variables x_1, x_2, x_3 and x_4 are respectively: $\lambda_{11} = 1/4$, $\lambda_{21} = 1/2$, $\lambda_{31} = 1/2$ and $\lambda_{41} = 1/2$. We assume that point-fix precision ϵ is equal to 0.005.

The equation system is the one given in the example 2. The different iterations of the resulting similarity computation are provided in Table 1.

Iterations	**0**	**1**	**2**	**3**	**4**
$x_1 = \max(0.68, x_2, x_3, \frac{1}{4} * x_4)$	0	0.68	0.9	0.9	0.9
$x_2 = \max(0.1, \frac{1}{2} * x_1)$	0	0.1	0.34	0.45	0.45
$x_3 = \max(0.9, \frac{1}{2} * x_1)$	0	0.9	0.9	0.9	0.9
$x_4 = \max(0.42, x_1)$	0	0.42	0.68	0.9	0.9

Table 1. *Example of iterative similarity computation*

The solution of the equation system is $X = (0.9, 0.45, 0.9, 0.9)$. This corresponds to the similarity scores of the four reference pairs. The fix-point has been reached after four iterations. If we fix the reconciliation threshold T_{rec} at 0.80, then we obtain three reconciliation decisions: two cities, two museums and two paintings.

1.3 Adaptations made for the evaluation

In order to perform the evaluation of LN2R system on the data sets provided by IM@OAEI2010 evaluation campaign we were faced to do some choices and adaptations. As LN2R system assume that the data sets conform to the same ontology, we have performed the following steps:

1. manual alignment of the two ontologies (schemas),
2. choose a federated ontology (one among the two considered ontologies) and
3. transform the other ontology to the chosen one.

Thanks to the small size of the considered ontologies and to their structural and semantic closeness, the above steps were performed easily. For example, for the restaurant data set, the difference between the two ontologies is that the $category$ and $city$ are object properties in one ontology and data properties in the other.

In addition to this, we have transformed the LN2R output to comply with the OAEI alignment format.

1.4 Link to the system and parameters file

The LN2R system (including the parameters file) can be downloaded at:
`http://www.lri.fr/~sais/IM-OAEI10/LN2RSystem.zip`.

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

The results that are obtained by LN2R in the instance matching track of OAEI 2010 campaign can be found at:
`http://www.lri.fr/~sais/IM-OAEI10/LN2RResults.zip`.

2 Results

In this section we present our comments on the results obtained from the first participation of LN2R in the Instance matching track IM@OAEI 2010. We have tested LN2R system on person and restaurant (PR) data sets.

To evaluate our system we have compared its results on the different data sets with the provided gold-standard and we have computed the recall, the precision and the F-measure.

2.1 Person1 and Person2 data sets

In this track, participants are asked to find all correct alignments between person' instances for the two data sets $person1$ and $person2$. Each data set contains the OWL ontologies, the two RDF files to be reconciled and the reference alignments (gold-standard) file.

Since, in LN2R, the reconciliation decisions are based on a reconciliation threshold, we have performed several tests by varying the threshold value from 0.6 to 1. The best results are obtained for a threshold of 0.75 for both data sets: (i) for $person1$ data set, we have obtained the maximum F-measure of 100 % for a recall of 100 % and a precision of 100% and (ii) for the $person2$ data set we have obtained a F-measure of 93% for a recall of 88.25 % and a precision of 99.4 %.

Furthermore, as our method is global in the sense that the reconciliation decisions between instances are propagated to other pairs of instances through the relations which link them together, we have also inferred alignments between address instances. Thanks to the functionnality of $has - address$ property, we have obtained 500 alignments between address instances for $person1$ data set and 355 alignments for $person2$ data set for a threshold of 0.75.

In addition to reconciliations (positive alignments), we also infer non reconciliations between instances, thanks to the reasoning on ontology knowledge, like disjunctions and UNA. In $person1$ and $person2$ ontologies we have declared the disjunction between $person$ and $address$ classes which leads to the inference of a non reconciliation between every person instance and every address instance. These non reconciliations are very useful in a filtering step where unnecessary comparisons and similarity computation are avoided.

2.2 Restaurant data set

In this track, participants are asked to find all correct alignments between restaurant' instances of the $restaurant1$ data set. It contains the OWL ontologies, the two RDF files to be reconciled and the reference alignments (gold-standard) file.

As we have done for the person data sets, we have also performed several tests of the system by varying the threshold value from 0.6 to 1. Comparing to the gold standard, the best results are obtained for a threshold of 0.85. We have obtained a F-measure of 75.3 % for a recall of 75 % and a precision of 75.67 %.

Similarly to the person data sets, we have also inferred a set of alignments between address instances thanks to the propagation mechanism. In a filtering step, we have also inferred a set of non reconciliation (negative alignments) between restaurant and address instances.

By analyzing the results of the restaurant data set, we have noticed some mistakes in the provided reference alignments: correct alignments are missed (see example 1) and some given alignments are wrong (see example 2).
Example 1: the two instances (http://www.okkam.org/oaie/restaurant1-Restaurant16, http://www.okkam.org/oaie/restaurant2-restaurant26) should be included in the gold-standard because they refer to the same restaurant. There descriptions are as follows:
(1) ['name: patina', 'category: californian', 'phone_number:213/467-1108' , has-address[street:'5955 melrose ave.', is_in_city[name:los angeles']]]
(2) ['name: patina', has_category['name:californian'], 'phone_number:213-467-1108' , has-address['city:los angeles', street:'5955 melrose ave.']]

Example 2: the two instances (http://www.okkam.org/oaie/restaurant1-Restaurant2, http://www.okkam.org/oaie/restaurant2-restaurant2) should be removed from the gold-standard because they do not refer to the same restaurant. There descriptions are as follows:
(1) ['name: hotel bel air', 'category: californian', 'phone_number: 310/472-1211' , has-address[street:'701 stone canyon rd.', is_in_city[name:bel air']]]
(2) ['name: art's deli', has_category['name:delis'], 'phone_number:818-762-1221' , has-address['city:studio city', street:'12224 ventura blvd.']]

3 General comments

3.1 Comments on the results

The main strength of our system is its capacity to ensure a good precision in the results. In the person data set, it shows its strength over ontology knowledge reasoning and similarity measures adaptation. LN2R system is also able to minimize the number of comparisons thanks to the filtering step which leads to improvement in running time. The weak points are: the absence of knowledge on the functionality of properties impacts the performance of the system. LN2R system works on data sets which should conform to the same ontology are for which the ontology alignment is already performed.

3.2 Discussions on the way to improve the proposed system

Our system may be improved by several ways. We are studying how LN2R can be extended to take into account alignments between classes and properties. We also want to optimize the system in order to insure its scalability.

3.3 Comments on the OAEI 2010 test cases

It will be interesting to provide test cases where the alignment inference is global. It means that the alignments may concern several kinds of entities e.g. persons, addresses, books, etc. It will be useful also to have data sets which conform to the same ontology or at least give the alignments between their corresponding ontologies.

4 Conclusion

Instance matching is very important to realize the semantic Web ambitions by facilitating interoperability of ontology based applications. In this paper, we have presented the promising results of LN2R system for its first participation in the instance matching track of OAEI 2010. By this experience, we have shown LN2R strengths when the ontology knowledge is rich. In the person data sets, LN2R has obtained very good results and reasonable ones for the restaurant data sets. As future work, we will study the extension of LN2R to the general problem of matching ontologies with instances. We also plan to optimize LN2R by designing a distributed inference algorithm.

References

1. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. (2005) 146–171
2. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal **10**(4) (2001) 334–350
3. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
4. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: A survey. IEEE Trans. on Knowl. and Data Eng. **19**(1) (2007) 1–16
5. Saïs, F., Pernelle, N., Rousset, M.C.: Combining a logical and a numerical method for data reconciliation. J. Data Semantics **12** (2009) 66–94
6. Dong, X., Halevy, A.Y., Madhavan, J.: Reference reconciliation in complex information spaces. In: SIGMOD Conference. (2005) 85–96
7. Saïs, F., Pernelle, N., Rousset, M.C.: L2r: A logical method for reference reconciliation. In: AAAI. (2007) 329–334
8. Golub, G.H., Loan, C.F.V.: Matrix computations (3rd ed.). Johns Hopkins University Press, Baltimore, MD, USA (1996)
9. Cohen, W.W., Ravikumar, P.D., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: IIWeb. (2003) 73–78