

TaxoMap alignment and refinement modules: Results for OAEI 2010

Fayçal Hamdi¹, Brigitte Safar¹, Nopal B. Niraula², and Chantal Reynaud¹

¹ LRI CNRS UMR 8623, Université Paris-Sud 11, Bat. G, INRIA Saclay
2-4 rue Jacques Monod, F-91893 Orsay, France
`firstname.lastname@lri.fr`

² The University of Memphis
Memphis, TN, USA
`nb.niraula@gmail.com`

Abstract. TaxoMap is an alignment tool which aims to discover rich correspondences between concepts (equivalence relations (*isEq*), subsumption relations (*isA*) and their inverse (*isMoreGnl*) or proximity relations (*isClose*)). It performs an oriented alignment (from a source to a target ontology) and takes into account labels and sub-class descriptions. This new implementation of TaxoMap uses a pattern-based approach implemented in the TaxoMap Framework helping an engineer to refine mappings to take into account specific conventions used in ontologies.

1 Introduction

TaxoMap was designed to retrieve useful alignments for information integration between different sources. The alignment process is then **oriented** from ontologies that describe external resources (named *source* ontology) to the ontology (named *target* ontology) of a web portal. The target ontology is supposed to be well-structured whereas source ontology can be a flat list of concepts.

TaxoMap makes the assumption that most semantic resources are based essentially on classification structures. This assumption is confirmed by large scale ontologies which contain rich lexical information and hierarchical specification without describing specific properties or instances. Then, to find mappings we use the following available elements: labels of concepts and hierarchical structures.

The new implementation of TaxoMap introduces a step of refinement of mappings (the alignment results) which extends the alignment process and completes it.

We take part to two tests. We hope the new step of refinement helps us to perform better in terms of precision of generated mappings.

2 Presentation of the System

2.1 State, Purpose and General Statement

TaxoMap has been designed to align owl ontologies $O = (C, H)$. C is a set of concepts characterized by a set of labels and H is a subsumption hierarchy which contains a set

of *isA* relationships between nodes corresponding to concepts. The alignment process is an oriented process which tries to connect the concepts of a source ontology O_S to the concepts of a target ontology O_T . The correspondences found are equivalence relations (*isEq*), subsumption relations (*isA*) and their inverse (*isMoreGnl*) or proximity relations (*isClose*).

To identify these correspondences, TaxoMap implements techniques which exploit the labels of the concepts and the subsumption links that connect the concepts in the hierarchy. The morpho-syntactic analysis tool, *TreeTagger* [1], is used to classify the words of the labels of the concepts and to divide them into two classes, *full words* and *complementary words*, according to their category and their position in the labels. At first the repartition between *full* and *complementary words* is used by a similarity measure that compares the tri-grams of the labels of the concepts [2] and gives more weight to the common *full words*. Then it is also used by the alignment techniques. For example, one technique named *LabelInclusion* generates an *isA* mapping between c_s and c_{tmax} if (1) the concept c_{tmax} is the concept of O_T having the highest similarity value with the concept c_s of O_S , (2) one of the labels of c_{tmax} is included in one of the labels of c_s , (3) all the words of the included label of c_{tmax} are classified as *full words* by *TreeTagger*.

Given a concept c_s of the ontology source O_S , our similarity measure identify the concept c_{tmax} of the target ontology O_T which have the highest similarity with c_s . The alignment techniques are then used to decide if the concept c_s can be effectively aligned with this concept c_{tmax} and which relation should be established between the two concepts, or whether, another concept of O_T must be chosen. A proposed mapping belongs to a single method, a concept of O_S can be aligned at most with one concept of O_T . In contrast, the concepts of O_T may be involved in several proposed alignments.

The main methods used to extract mappings between a concept c_s in O_S and a concept c_t in O_T are:

- Label equivalence: An equivalence relationship, *isEq*, is generated if the similarity between one label of c_{tmax} and one label of c_s is greater than a threshold (Equiv.threshold).
- Label inclusion (and its inverse): If one of the labels of c_{tmax} is included in one of the labels of c_s , and if all words of included label are full words, we propose a subclass relationships $\langle c_s \text{ isA } c_{tmax} \rangle$. Inversely, if one of the labels of c_s is included in one of the labels of c_{tmax} , we propose the relationships $\langle c_s \text{ isMoreGnl } c_{tmax} \rangle$.
- High lexical similarity: If the similarity measure of c_{tmax} is greater than a threshold (HighSim.threshold) and if one of its labels shares at least two full words in common with one of the labels of c_s , without being including in the labels of c_s , the heuristic generates the relationship $\langle c_s \text{ isClose } c_{tmax} \rangle$.
- Reasoning on similarity values: Let c_{tmax} and c_{t2} be the two concepts in O_T with the highest similarity measure with c_s , the relative similarity is the ratio of c_{t2} similarity on similarity c_{tmax} . If the relative similarity is lower than a threshold (isA.threshold), one of the two following techniques can be used:

- the relationship $\langle c_s \text{ isClose } c_{tmax} \rangle$ is generated if the similarity of c_{tmax} is greater than a threshold (`isClose.thresholdMax`).
 - an isA relationship is generated between c_s and the father of c_{tmax} if the similarity of c_{tmax} is greater than a second threshold (`isA.thresholdMax`).
- Best similarity: If none of the above techniques is applicable, the relationship $\langle c_s \text{ isClose } c_{tmax} \rangle$ is generated if the similarity of c_{tmax} is greater than a threshold (`Better.thresholdMax`).

Mappings identified by TaxoMap are generated in the Alignment format used as a standard in the OAEI campaign. We added to this format the information about the names of the techniques that generated mappings. The aim is to facilitate the specification of treatments exploiting the mappings generated by those techniques. All these pieces of information are stored in a relational mappings database which can then be queried using *SQL* queries. This allows, in particular, to present the generated mappings to the expert in the validation phase, technique by technique.

In the OAEI 2010 campaign, only equivalence relations will be evaluated in the alignment contest. This has two important implications on our results:

1) In fact, all mappings generated by the label inclusion techniques that lead to a subsumption relation isA are wrong if they are converted into equivalence relations. We do not use them here.

2) a concept of O_T must have only one equivalent concept in O_S , so if we consider the mappings leading to the proximity relations, all mappings which connect a concept of O_S to a concept of O_T which is already involved in an equivalence relations are also false.

We will see in the next section how the TaxoMap refinement module [8] will allow us to remove these incorrect mappings.

2.2 The Mapping Refinement Workflow

We proposed an environment allowing to specify and perform refinement treatments applied on the prior obtained mappings. At first, this environment will be used to improve the quality of an alignment provided by TaxoMap. Subsequently, it will be used for other treatments based on mappings as enriching, restructuring or merging ontologies.

An important feature of the approach is to allow a declarative specification of treatments based on particular alignment results, concerning particular ontologies and using a predefined vocabulary. Treatments which can be specified depend on the characteristics of the concerned ontologies and on the task to be performed (at first mapping refinement and subsequently ontology merging, restructuring, enriching). These treatments are thus associated to independent specification modules, one for each task, each having their own vocabulary. The approach is extensible and a priori applicable to any treatment based on alignment results.

We present the Mapping Refinement Pattern Language (MRPL) used to specify mapping refinement pattern. This language differs from the one defined in [4] especially because it includes patterns which test the existence of mappings generated by

alignment techniques.

The vocabulary of MRPL contains:

- *a set of predicate constants.* We distinguish three categories of predicate constants: the predicate constants relating to the type of techniques applied in the identification of a mapping by TaxoMap, the predicate constants expressing structural relations between concepts of a same ontology, the predicate constants expressing terminological relations between labels of concepts.
- *a set of individual constants:* $\{a, b, c, \dots\}$
- *a set of variables:* $\{x, y, z, \dots, _ \}$ where $_$ is an unnamed variable used to represent parameters which do not need to be precised.
- *a set of built-in predicates:* $\{Add_Mapping, Delete_Mapping\}$
- *a set of logical symbols:* $\{\exists, \wedge, \neg\}$

MRPL allows the definition of a **context part** which must be satisfied to make the execution of a pattern possible, and of a **solution part** which expresses the process to achieve when the **context part** is satisfied. The **context part** is a logical formula where: *Variables and constants are terms, if α and β are terms and P is a predicate symbol with two places then $P(\alpha, \beta)$ is a formula, if α, β and γ are terms and P is a predicate symbol with three places then $P(\alpha, \beta, \gamma)$ is a formula, if ϕ and ψ are formulae then $[\phi \wedge \psi]$ is a formula, if ϕ is a formula then $[\neg \phi]$ is a formula, if ϕ is a formula and v is a variable then $\exists v\phi$ is a formula.*

Context part of pattern

The **context part** tests (1) the technique used to identify the considered mapping, (2) the structural constraints on mapped elements, for example, the fact that they are related by a subsumption relation to concepts verifying or not some properties, or (3) the terminological constraints, for example, the fact that the labels of a concept are included in the labels of other concepts. These conditions are represented using formulae built from predicate symbols. So, we distinguish three kinds of formula according to the kind of predicate symbols used.

The formulae related to the type of techniques applied in the identification of a mapping by TaxoMap. By testing the existence in the mappings database of a particular relation generated by a given technique, we build formulae that implicitly test the conditions for the application of this technique. For example the formula *isAStrictInclusion*(x, y) tests the existence of a mapping *isA* generated between two concepts x and y using the technique searching *LabelInclusion*, t_2 . It validates implicitly at the same time all the conditions for the application of t_2 , i.e. (1) the concept y is the concept of O_T having the highest similarity value with the concept x of O_S , (2) one of the labels of y is included in one of the labels of x , and (3) all the words of the labels of y are classified as *full words* by *TreeTagger*. TaxoMap includes several alignment techniques. Thus, several predicate symbols leading to formulae of that kind are needed. More formally, let:

$R_M = \{isEq, isA, isMoreGnl, isClose\}$, the set of correspondence relations used by TaxoMap,

$T = \{t_1, t_2, \dots\}$, the set of techniques.

T_M , the table storing generated mappings in the form of 4-tuple (x, y, r, t) where $x \in C_S, y \in C_T, r \in R_M, t \in T$. The pairs of variables (x, y) which can instantiate these formulae will take their values in the set $(x, y) \mid (x, y, r, t) \in T_M$. The predicate symbols necessary for the task of refinement presented in this paper are *isEquivalent* and *isAStrictInclusion* the semantics of which are the following:

- *isEquivalent* (x, y) is true iff $\exists(x, y, isEq, t_1) \in T_M$
- *isAStrictInclusion* (x, y) is true iff $\exists(x, y, isA, t_2) \in T_M$
- *mapping* (x, y) is true iff $\exists(x, y, -, -) \in T_M$

The formulae expressing structural relations between concepts x and y of the same ontology $O = (C, H)$. Since the aim of TaxoMap is the alignment of taxonomies, the structural relations considered here are subsumption relations. If the approach was used with another alignment tool, other relations could be considered. Note that the instances of variables in these formulae will be constrained, either directly because they instantiate the previous formulae, related to the type of the applied techniques, or indirectly by having to be in relation with other instances.

- *isSubClassOf* (x, y, O) is true $\Leftrightarrow isA(x, y) \in H$
- *isParentOf* (x, y, O) is true $\Leftrightarrow isA(y, x) \in H$

The formulae expressing terminological relations between the labels of the concepts: not detailed here because not used in the examples of this paper.

Other formulae as - *conceptsDifferent* (x, y) is true $\Leftrightarrow ID(x) \neq ID(y)$ with $ID(x)$ is the identifier of the concept x .

Solution part of pattern

A **context part** is associated to a **solution part** which is a set of actions to be performed. This set of actions is modeled by a conjunction of built-in predicates executed in a database. The built-in predicates are defined as follows:

- *Add_Mapping* (x, y, r) has the effect of adding a tuple to the table T_M which becomes $T_M \cup \{(x, y, r, t)\}$ where r and t are fixed in the treatment condition by instantiating the predicate corresponding to the type of technique associated with the considered mapping.
- *Delete_Mapping* $(x, y, -)$ has the effect of removing a tuple from the table T_M which becomes $T_M - \{(x, y, -, -)\}$.

Mapping Refinement Pattern used in OAEI

Pattern-1: This pattern concerns mappings generated by the technique t_1 , connecting a concept y of the target ontology O_T with a concept x of the source ontology O_S . Because a concept y of the target ontology O_T must be involved in at most one equivalence relation, mappings involving y and obtained from other techniques than t_1 should

be removed.

Context part of Pattern-1:

$\exists x \exists y (isEquivalent(x, y) \wedge \exists z (mapping(z, y) \wedge conceptDifferent(z, x)))$

Solution part of Pattern-1:

Delete_Mapping(z, y, -)

For the anatomy subtask 4, if we know a set of reference mappings, we could express a new refinement pattern to remove generated mappings that relies a concept of the target ontology (y) to different concepts of the source ontology ($x, z, ..$).

We should define the new predicate *referenceMapping(x, y)* as follow: *referenceMapping(x, y)* is true iff $\exists (x, y) \in Reference_Mapping$.

Context part of Pattern-2:

$\exists x \exists y (referenceMapping(x, y) \wedge \exists z (mapping(z, y) \wedge conceptDifferent(z, x)))$

Solution part of Pattern-2:

Delete_Mapping(z, y, -)

2.3 Adaptations made for the Evaluation

2.4 Link to the system and parameters file

TaxoMap requires:

- Java (Version 1.5 and above)³

The version of TaxoMap (with parameter files) used in 2010 contest can be downloaded from:

- <http://www.lri.fr/~hamdi/TaxoMap/TaxoMap.html>

2.5 Link to the Set of Provided Alignments

The alignments produced by TaxoMap are available at the following URLs:

<http://www.lri.fr/~hamdi/OAEI10/anatomy/>

<http://www.lri.fr/~hamdi/OAEI10/directory/>

³ <http://java.sun.com>

3 Results

3.1 Anatomy Test

The anatomy real world case is to match the Adult Mouse Anatomy (denoted by *Mouse*) and the NCI Thesaurus describing the human anatomy (tagged as *Human*). *Mouse* has 2,744 classes, while *Human* has 3,304 classes. We considered *Human* as the target ontology as is it well structured and larger than *Mouse*.

TaxoMap performs the alignment in about 12 minutes.

As only equivalence relationships will be evaluated in the alignment contest, we did not use this year the techniques which generate *isA* relationship (except in the Task 3) and we change *isClose* mapping to equivalence. In addition, we use the refinement pattern described above to delete mappings between a concept of the target ontology that was already aligned with an equivalence mapping. As a result, we found fewer mappings than last year but we hope that the precision will be better than the results of the last year [5].

3.2 Directory Test

The directory task consists of Web sites directories like Google, Yahoo! or Looksmart. To date, it includes 4,639 tests represented by pairs of OWL ontologies. TaxoMap takes about 40 minutes to complete all the tests.

4 General Comments

4.1 Results

The new version of TaxoMap improves significantly the results on the previous version of TaxoMap in terms of runtime and precision of generated mappings. The new implementation offers extensibility and modularity of code. TaxoMap can be parameterized by the language used in ontologies, the choice of used techniques and different thresholds.

4.2 Future Improvements

The following improvements can be made to obtain better results:

- To take into account all concepts properties instead of only the hierarchical ones.
- To use WordNet as a dictionary of synonymy. The synsets can enrich the terminological alignment process if an *a priori* disambiguation is made.
- To develop the remaining structural techniques which proved to be efficient in last experiments [6] [7].

5 Conclusion

This paper reports our participation to OAEI campaign with the new implementation of TaxoMap. Our algorithm proposes an oriented mapping between concepts. Our participation in the campaign allows us to test the robustness of TaxoMap and new structural techniques.

References

- [1] Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees, International Conference on New Methods in Language Processing (1994)
- [2] Lin, D. : An Information-Theoretic Definition of Similarity. ICML. Madison. (1998) 296–304
- [3] Maedche, A. and Staab S. Measuring Similarity between Ontologies, EKAW (2002)
- [4] Scharffe, F.: Correspondence Patterns Representations. PhD thesis, University of Innsbruck, 2009.
- [5] Hamdi, F., Safar, B., Nabal B. Niraula and Reynaud, C. TaxoMap in the OAEI 2009 alignment contest, Proceedings of the ISWC'09 Workshop on Ontology Matching OM-09 (2009)
- [6] Reynaud, C. and Safar, B. When usual structural alignment techniques don't apply, The ISWC'06 Workshop on Ontology matching (OM-06), (2006)
- [7] Reynaud, C. and Safar, B. Exploiting WordNet as Background Knowledge, The ISWC'07 Workshop on Ontology Matching (OM-07), (2007)
- [8] Hamdi, F., Reynaud, C. and Safar, B., Pattern-based Mapping Refinement, 17th International Conference on Knowledge Engineering and Knowledge Management, EKAW (2010)