TaxoMap in the OAEI 2007 alignment contest

Haïfa Zargayouna, Brigitte Safar, and Chantal Reynaud

LRI, Université Paris-Sud 11, Bât. G, INRIA Futurs
2-4 rue Jacques Monod, F-91893 Orsay, France
firstname.lastname@lri.fr

Abstract. This paper presents our first participation in the OAEI 2007 campaign. It describes an approach to align taxonomies which relies on terminological and structural techniques applied sequentially. We performed our method with various taxonomies using our prototype, TaxoMap. Previous experimental results were encouraging and demonstrate the relevance of this alignment approach. In this paper, we evaluate the strengths and weaknesses of TaxoMap in the context of the OAEI campaign where the ontologies to align are different from taxonomies we are used to deal with.

1 TaxoMap: Context

The increasing amount of information sources available on the Web requires techniques providing integration. Ontologies define concepts relative to particular application domains. They have become central in information integration because they allow description of content of integrated sources and make the vocabulary to be used in the queries explicit. The ontology alignment task (correspondences or mappings finding) is particularly important in information integration systems because it allows several heterogeneous systems which have their own ontology to be used jointly.

Our alignment system, TaxoMap, has been designed in the setting of query answering in the food risk domain. We aimed at increasing answers delivered by a web portal thanks to information provided by others sources annotated by semantic resources. Querying the portal was supported by a global schema exploited by a query interface which had to be reused without any change.

TaxoMap was designed to discover alignments between this global and rich schema and much simpler semantic resources of other sources. The experts required that the retrieval process should not be altered by the alignment process.

The alignment process is then **oriented** from an ontology, named source ontology, (for instance an ontology associated to external resources) to a target ontology (for instance the ontology of a web portal).

Moreover, our mapping techniques are strict: only concepts that have strictly the same label are matched with an equivalence relation. The remaining concepts of the source ontology are matched with a subclass relation which denotes a proximity relation. Therefore, TaxoMap proposes essentially subclass relation mappings.

We assume that often, content of information sources is not specified a lot. Simple ontologies reduced to classification structures, i.e. taxonomies, are the only way to describe their content. Moreover, we suppose that the taxonomies that we align are heterogeneous, describing the same domain in different vocabularies and structures, the target taxonomy being well-structured whereas the source taxonomy perhaps not. In this context, the approaches that rely on OWL data representations, exploiting all the ontology language features, do not apply. To find mappings, we can only use the following available elements: the labels of concepts in both taxonomies and the structure of the target taxonomy.

We propose several alignment techniques whose aim is to discover classes of mappings between taxonomies belonging to a general methodology usable across different application areas. Classes of mappings are categorized into probable mapping and potential mapping classes (i.e. to be confirmed or refuted manually). The mapping process can be viewed as an execution of various techniques invoked in sequence, namely terminological followed by structural techniques.

Terminological techniques are based on string comparisons. They discover mappings that exploit the richness of the labels of the concepts. These techniques are efficient in the sense that they provide high-quality alignments corresponding to probable mappings. Unfortunately, they are not sufficient. Many of the mappings are undiscovered. So, we extend the terminological techniques with structural ones.

The paper is organized as follows. Section 2 describes the alignment approach and the adaptations made for the evaluation. In section 3, we present the results of the experiments we have done so far.

2 Presentation of the system

2.1 State, Purpose and General Statement

The objective of our approach is to generate mappings between taxonomies. For us, a taxonomy T is a pair (C, H_C) consisting of a set of concepts C arranged in a subsumption hierarchy H_C. A concept c is only defined by two elements: a label and subclass relationships. The label is a name (a string) that describes entities in natural language and which can be an expression composed of several words. A subclass relationship establishes links with other concepts. It is the only semantic association used in the classification. A taxonomy is generally represented by an acyclic graph where concepts are represented by nodes connected by directed edges corresponding to subclass links.

The objective is to map the concepts of the source taxonomy T_{Source} to the concepts of the target taxonomy T_{Target}. It is an **oriented process** from T_{Source} to T_{Target}. Hence, we define the mapping as follows: Given two taxonomies, T_{Source} and T_{Target}, mapping T_{Source} with T_{Target} means that: for each concept (node) c_S in T_{Source}, we try to find a corresponding concept (node), c_T in T_{Target}, linked to c_S with an equivalence or a subclass relation.

The alignment process aims at finding one-to-one mappings between single concepts and establishing two types of relationships, equivalence and subclass relationships defined as follows.

Equivalence relationships An equivalence relationship, *isEq*, is a link between a concept in T_{Source} and a concept in T_{Target} with labels assumed to be similar.

Subclass relationships Subclass relationships are usual *is-A* class links. When a concept c_S of T_{Source} is linked to a concept c_T of T_{Target} with such a relationship, c_T is be considered as a super concept of c_S.

2.2 Techniques Used

All our alignment techniques are based on Lin's similarity measure ($Sim_{LinLike}$)[1] computed between each concept c_S in T_{Source} and all the concepts in T_{Target}. This measure compares strings and has been adapted to take into account the importance of words inside expressions. Terminological and structural techniques are used (see figure 1) and are applied in sequence to maximize the efficiency of the overall alignment process. For each technique, the objective is to select the best concept in T_{Target} among many mapping candidates MC (with a similarity measure not being null), the best concept having not necessarily the highest similarity measure.

Terminological techniques are executed first. Being based on the richness of the labels of the concepts, they provide the most relevant mappings. They are performed in three steps:

- **Search for equivalents** The first relationships to be discovered are equivalence relationships, which map concepts with a similarity measure corresponding to a strong similarity (greater than a threshold which has been set to 1 in our experiments).
- **Labels inclusion** We consider inclusion of name strings for which we propose a subclass mapping between c_S and c_T if c_T is the concept in T_{Target} with the highest similarity measure and if the name string of c_T is included in the name string of c_S in T_{Source}.
- **Relative similarity** If the name string of the concept c_T of T_{Target} with the higher similarity measure is not included in the name string of c_S, but if its similarity measure is significantly highest than the measure of the others, c_T is considered as a brother of c_S and the system proposes a subclass relationship between c_S and the father node of c_T.

All of the above techniques are performed in sequence. They merely rely on the values of similarity measures and lead to mappings which are generally reliable but not always sufficient in number.

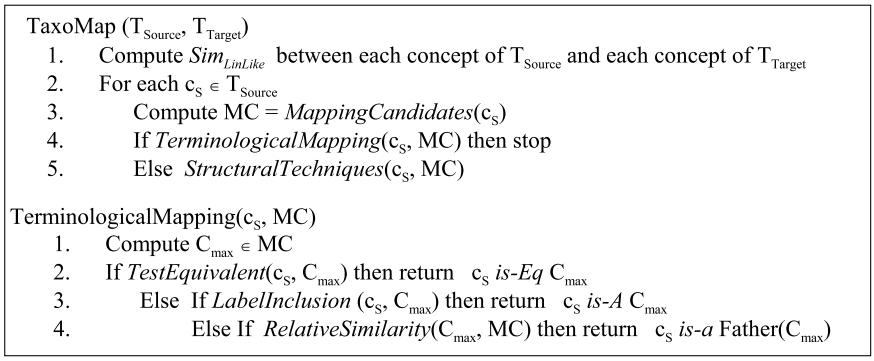

Fig. 1. The alignment process

When terminological techniques are not sufficient, complementary techniques are used to provide additional mappings. In this regard, we propose the following complementary techniques:

- Take advantage of structural features in the target taxonomy,
- Exploit the hierarchical structure of additional background knowledge,
- Deduce new mappings from prior defined ones.

We do not detail these techniques as they were not applied in the experiments we report below. The interested reader can refer to [3, 2] for a detailed presentation of these techniques.

2.3 Adaptations made for the Evaluation

The TaxoMap prototype is written in Java and takes as input two taxonomies whose format is compliant with that of TaxoMap[1]. TaxoMap outputs a file per technique used (equivalence, inclusion, proximity, etc.). We have developed two conversion modules to link our application to the API Alignment. They are:

- *OWL2TM:* It parses ontologies in OWL or RDF and generates taxonomies in TaxoMap format where only labels and subclass relationships are taken into account. This module generates a table of correspondences which stores URIs with their associated labels.
- *TM2Align:* It generates an output from TaxoMap internal structure, having the alignment output format (RDF/XML) which regroups all the information stored in TaxoMap's output.

[1] A text format in which concepts are presented in their context (below their super concept) and defined by their label and their level in the hierarchy.

2.4 Link to the Set of Provided Alignments

The alignments produced by TaxoMap are available at the following URLs:
http://www.lri.fr/~haifa/benchmarks/
http://www.lri.fr/~haifa/anatomy/

3 Results

3.1 Benchmark Tests

Tests 101-104 Since our algorithm only considers labels and subclass relations and only provides mappings for concepts, the recall is low even for the reference alignment (#101-#101).

Test	Precision	Recall	F-Measure
101-101	1.00	0.34	0.5
102-101	NaN	NaN	NaN
103-101	1.00	0.34	0.5
104-101	1.00	0.34	0.5

Table 1. Results from 101 to 104

Tests 201-266 Given below are some results concerning alterations on labels and hierarchies. Alterations on properties, instances and comments have no effect on our algorithm since it ignores these descriptions.

Our algorithm relies on labels of concepts, so tests where labels were suppressed or replaced by random string or translated[2] have produced no mapping.

In most remaining tests, the precision was very high. This is encouraging since our main objective is to increase ontology mapping precision.

We considered the reference ontology #101 as the Target ontology to fit to our initial hypotheses (see section 2.1). This restriction distorts the interpretation of the results. In fact, our alignment is oriented and affects the results. We generate alignments from concepts of #X to concepts of #101 (i.e (ci_X *isEq* cj_{101}), (cn_X *isEq* cj_{101}), etc.). However reference alignments contain #101 as the first ontology to align. So to be comparable with these alignments, we changed the order of the results (cj_{101} *isEq* ci_X), (cj_{101} *isEq* cn_X), etc.). This leads to generate 1:n relations and explains the cases where the precision has deteriorated slightly.

[2] For official runs, we only take into account English labels. However, when this restriction is not applied, we obtain few alignments because there could exist some common roots between labels even if the language is different (it is the case between French and English).

Tests 301-304 Tests with real ontologies presented coherent results with the rest of tests. The precision is high, but the recall is low because alignments concern also properties which are not taken into account by our algorithm.

Test	Precision	Recall	F-Measure
301-101	1.00	0.21	0.35
302-101	1.00	0.21	0.35
303-101	0.80	0.24	0.37
304-101	0.92	0.34	0.5

Table 2. Results from 301 to 304

3.2 Anatomy Test

We considered *nci_anatomy* as the target taxonomy as is it well structured and larger than *mouse_anatomy*. With respect to the chosen anatomy, we were only able to apply the terminological techniques due to the large size of the taxonomies.

In order to remedy the above problem, we attempted to reduce the size of taxonomies by performing a filtering phase consisting of detecting disjoint domains between the taxonomies and deleting sub-hierarchies that had no match in the terminological mapping phase. The filtering phase did not detect disjoint subcategories, common concepts between the two taxonomies are homogeneously distributed in all sub-hierarchies of T_{Target}. The two taxonomies seem to be homogeneous and no subsequent part can be deleted without deteriorating the performances of the system.

The reference alignment contains only equivalence correspondences between concepts of the ontologies. In order to have a little insight on the relevance of the extracted relations, we transformed *is-A* relations into equivalence ones with a confidence value of 0.5. This transformation seems adequate for some cases, where labels are slightly different and where the label of a concept c_S in *mouse_anatomy* is included in the label of a concept c_T in *nci_anatomy*. For example, the concepts *abducens VI nerve* and *abducens nerve* are identified as not equivalent but considered as being related by an is-A relation (*abducens VI nerve is-A abducens nerve*). When we transform this relation into an equivalence one (*abducens VI nerve isEq abducens nerve*), the alignment remains good.

This, however, is not always possible and can lead to misinterpretation of our results. It is especially serious when the subsumption relation is proposed between c_S and a father of c_T. For example, the relative similarity technique discovers a similarity between *cervicothoracic ganglion* and *thoracic ganglion*, and then infers subsumption relation between *cervicothoracic ganglion* and *sympathetic ganglion*[3] (*cervicothoracic ganglion is-A sympathetic ganglion*). This re-

[3] *sympathetic ganglion* is the father of *thoracic ganglion* in T_{Target}.

lation seems to be adequate, but its transformation into an equivalence relation even with a low confidence seems to be nonsense.

So, for us the absence of is-A relations in the reference mappings is unsatisfactory. The labels of concepts are almost complex terms which are composed of terms of their super-concepts. Therefore, terminological techniques seem to fit to this sort of ontologies, particularly the label inclusion technique which find subsumption relaions.

We applied a strict threshold for the two runs we submitted[4]. Equivalence relations are found between concepts when labels are strictly the same. If there is some variation, the concepts are considered to be linked by a subsumption relation. We identified 941 equivalence mappings and 879 subsumption correspondences, but almost 900 concepts were left unmapped. Among these 900 unmapped concepts, 581 have a label which includes labels from other concepts and so can be candidates for an is-A mapping with a more relaxed threshold.

Hence, our first effort will be to solve the problem of large-scale ontologies in order to test the whole proposed techniques.

4 General Comments

4.1 Strengths and Weaknesses of the Results

TaxoMap was designed for semantic resources with poor concept descriptions. The benchmark tests were not adequate for testing the robustness of the terminological mappings as the values of the recall are influenced by property mappings and the reference alignment is oriented.

The anatomy test proposes interesting taxonomies. The generated mappings seem interesting. However, the results will, again, depend on the reference alignment. As subsumption relations are not evaluated, we should have relaxed the similarity threshold.

4.2 Ways to Improve the Proposed System

Our algorithm does not take into account properties and instances and only generates mappings between concepts, all of which seems to handicap our system. However, we believe that properties and instances in ontologies can be valuable when the ontologies are constructed rigorously. But if we consider the near future of semantic web where anyone can place his ontology on the web, it can lead to numerous light ontologies with only labels and subsumption relations.

The following improvements can be made to obtain better results:

- Take into account multi-label concepts in the terminological mappings.
- Exploit comments, if present, to enforce the confidence on the extracted mapping classes.

The main issue remaining concerns the way to process large-scale ontologies. This needs techniques to split ontologies and aggregate the returned results.

[4] The first run returns only equivalent mappings, the second returns equivalent and subsumption correspondences.

4.3 Comments on the OAEI measures

The precision and recall measures are necessary to have a general idea of the alignment performance. Nonetheless, they need to be adapted to the context of alignment where:

- All mappings have different weights. In fact, some mappings are more difficult to find and this difficulty should be quantified.
- Subsumption relations are probably less interesting than equivalence but are important in certain contexts (for instance query expansion).
- There should be a difference between a false alignment and an approximate one. The recall and precision measures take into consideration binary relevance (a mapping is considered as correct or not). They can be adapted in order to take into account graduate relevance (0 and 1 remain as "totally irrelevant" and "totally relevant" respectively, and intermediate values are assumed with various degrees of "partial relevance").

5 Conclusion

This paper reports our first participation in OAEI campaign. Our algorithm proposes an oriented mapping between concepts. This specificity leads to a misinterpretation of our results. The assessments of alignments consider only equivalent relations so we did not have an insight on the relevance of subclass relations. Despite these difficulties, our participation in the campaign opens perspectives to ameliorate our system.

References

[1] Lin, D. : An Information-Theoretic Definition of Similarity. ICML. Madison. (1998) p.296–304
[2] Reynaud, C., Safar, B. Exploiting WordNet as Background Knowledge. The ISWC'07 Ontology Matching (OM-07). (2007) to-appear
[3] Reynaud, C., Safar, B. When usual structural alignment techniques don't apply. The ISWC'06 workshop on Ontology matching (OM-06). Athens. (2006)