When usual structural alignment techniques don't apply

Chantal Reynaud[1] and Brigitte Safar[1]

University of Paris Sud-CNRS (LRI), INRIA (Futurs),
LRI, Building 490, 91405 Orsay Cedex, France
{chantal.reynaud, brigitte.safar}@lri.fr
http://www.lri.fr/~cr

Abstract. This paper deals with taxonomy alignment. It presents structural techniques of an alignment method suitable with a dissymmetry in the structure of the mapped taxonomies. The aim is to allow a uniform access to documents belonging to a same application domain, assuming retrieval of documents is supported by taxonomies.

1 Introduction

Our work focuses on taxonomy alignment techniques. Indeed, we assume that the description of the content of most todays information systems is often not very much specified and is based on very simple ontologies reduced to classification structures, i.e. taxonomies. Moreover, we suppose that the structures of the taxonomies that have to be aligned are heterogeneous and dissymmetric, one taxonomy being deep whereas the other one is flat. In this context, the approaches which relied on OWL data representations exploiting all the ontology language features don't apply. Similarity of two entities cannot be identified based on their similar properties or on the status of their respective parents and siblings, because these data are not available. We can only use the following available data: labels of concepts in both taxonomies, the structure of the deeper taxonomy and external linguistic resources such as WordNet.

The contribution of this paper is a mapping process composed of a sequence of various techniques designed to make best use of the characteristics of the taxonomies: very specialized taxonomies with only sub-class links, concepts with labels which are expressions composed of a lot of words, words common to a lot of labels. We classify the found mappings into two groups according to their relevance: probable mappings and potential mappings to be confirmed. The mapping process is generic, usable across application areas. It has been evaluated on real-world taxonomies and on test taxonomies extracted from a repository about ontology matching [6]. Experiments showed that the proposed techniques give very relevant mappings when the aligned taxonomies have the same characteristics as the taxonomies having motivated our approach.

2 The alignment approach

The objective of our approach is to generate mappings between taxonomies. For us, a taxonomy is a pair (C, H_C) consisting of a set of concepts C arranged in a subsumption hierarchy H_C. A concept is only defined by two elements: a label and subclass relations. The label is a string which can be an expression composed of several words. Subclass relations establish links with other concepts. It is the single semantic association used in the hierarchy. A taxonomy is generally represented by an acyclic graph where concepts are nodes connected by directed edges corresponding to subclass links. Given two structurally dissymmetric taxonomies, the objective is to map the concepts of the less structured one, the source taxonomy T_S, with concepts of the more structured one, the target taxonomy T_T. The alignment process is oriented from T_S to T_T. The goal is to find one-to-one mappings. Relations can be of two kinds: equivalence (*isEq*) and subclass (*isA*). So, for each concept c_S in T_S, we try to find a corresponding concept, c_T in T_T, linked to c_S with an equivalence or a subclass relation.

Alignment is based on the *Lin* similarity measure [1] computed between each concept c_S in T_S and all the concepts in T_T. This measure compares the tri-grams of the labels and has been adapted to take into account the importance of words inside expressions. From the measurements we compute MC, the set of mapping candidates of a concept c_S in T_S. MC includes concepts of T_T which have a high similarity value with c_S (only the three most similar concepts b_1, b_2, b_3 are retained) and Inc, the set of concepts of T_T with a label included in the label of c_S. Various techniques (terminological and structural cf.Fig.1) are then applied in sequence to select the most relevant concept among all the mapping candidates [3]. We are going to show that the most relevant concept is not necessarily the one with the highest similarity measure. Terminological techniques are executed first. In default of place, they will not be detailed here. These techniques lead to mappings which are generally reliable but not always sufficiently numerous. Therefore, they are completed by the structural mappings described in the next section. These latter techniques define a mapping as a correspondence between close concepts. If the suggested mapping from c_S to c_T is wrong, then the right mapping will be a relation from c_S to c'_T, with c'_T close to c_T in the taxonomy. It is a guide for the user who will not have to browse the whole target taxonomy when studying the results of the system.

```
$TaxoMap(T_S, T_T)$
For each $c_S \in T_S$ do
    For each $c_T \in T_T$ do $Sim_{LinLike}(c_S, c_T)$
    $MC \leftarrow$ MappingCandidates($c_S$)
    If TerminologicalTechniques($c_S, MC$) then stop
    Else StructuralTechniques($c_S, MC$)
```

Fig. 1. The Alignment process

3 Exploiting structural features

The two techniques presented in this section are structure based techniques leading to the discovery of subclass mappings. The first technique is performed on T_T whose structure is supposed to be the deepest. Then we use *WordNet* [2], exploiting its structure and its semantic relations.

3.1 Exploiting the structure of the target taxonomy: STR_T

This technique, denoted STR_T, works on MC, the set of mapping candidates of a concept c_S in T_S. The idea is to exploit the location of the elements of MC in T_T. Their proximity in the graph is considered to be a semantic proximity. We therefore try to identify the sub-graph rooted in a node associated to a concept which is not too general and such that this sub-graph groups a maximum of nodes of MC. It will represent a relevant context shared by most of the mapping candidates. We then consider that the involved concept c_S may be mapped with a node of this sub-graph. STR_T relies on the computation of the Lowest Common Ancestor, LCA, of a set of nodes in a graph, which is the node of greatest depth which is an ancestor of all the nodes of the set. Our goal is to find a LCA of the elements of MC which is not too high in the taxonomy. However the LCA node of a set of elements is all quite high in the graph since the elements are very distant from each other. We propose a measure, the relative density (DR), to evaluate sub-graphs grouping nodes of a sub-set of MC. For each sub-graph rooted in Anc, the LCA node, and grouping MC_{Anc} nodes, we compute $DR(Anc)$.

$DR(Anc)$ relies on three criteria: (1) the number of elements in MC_{Anc}, (2) Sim_{Lin_Like}, the similarity between the elements in MC_{Anc} and c_S,(3) the distance as the number of edges on the paths from each element of MC_{Anc} to Anc.

$$DR(Anc) = \frac{|MC_{Anc}| * \sum_{C_T \in MC_{Anc}} \mathrm{Sim}_{\text{LIN-Like}}(c_S, c_T)}{|MC| * \sum_{C_T \in MC_{Anc}} \mathrm{dist}(c_T, Anc)}$$

The sub-graph rooted in the Anc with the highest DR is considered to be the most relevant. C_{MaxAnc}, the node of this sub-graph with the highest similarity measure, will be the candidate selected for the mapping. If it belongs to Inc, the set of concepts with a label included into the label of c_S, it is suggested as a possible parent of the involved concept c_S. Otherwise, C_{MaxAnc} is proposed as a possible sibling and its parent (not necessarily Anc) will be suggested as a possible parent of c_S. As an example, Fig. 2 represents the sub-graph of T_T grouping the elements of $MC = \{b_1, b_2, b_3\} \cup Inc = \{$*beef*$\}$ for $c_S =$ *beef adipose tissue.* The node *Fresh meat* is the LCA for all the elements of MC with a distance of 7. However, *beef* is the LCA of three mapping candidates $\{$*beef*, b_2, $b_3\}$ with a distance of only 2. DR(*beef*) is the highest (cf.Fig.2). *beef connective tissue* is the node of this sub-graph with the highest similarity value to c_S. So *beef adipose tissue* will be a sibling of *beef connective tissue* and linked to *beef* with a

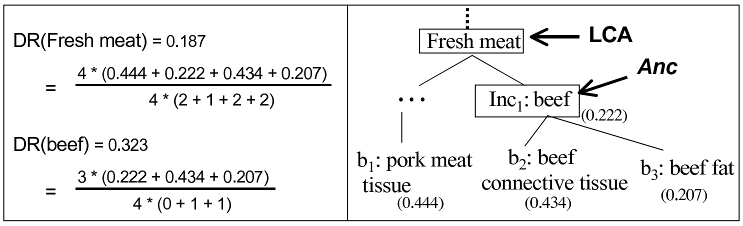

Fig. 2. Common ancestors and relative density

subclass relation. Note that this technique avoids mappings with concepts with a little higher similarity measure but meaningful in a context different from the one common to most of the MC (as b_1 in the example).

3.2 Exploiting the structure of *WordNet*: STR_W

The techniques seen up till now are not enough if the concepts are similar semantically but not syntactically. So, at that point, we propose to run STR_W. STR_W relies on the *hyperonymy/hyponymy* $WordNet$ structure to find the concept of T_T semantically similar to each concept of T_S not yet mapped. STR_W will be able to map, for example, *cantaloupe* with *watermelon* which are not synonyms but two specializations of *melon.*

Running STR_W assumes that the application root node, denoted $root_A$, has already been identified. It is the most specialized concept in $WordNet$ which generalizes all the concepts contained in the involved application domain. STR_W searches WordNet for the hypernyms of each term of T_S not yet mapped and of each term of T_T (according to all their senses) until $root_A$ is reached. For example, the result of a search on *cantaloupe* is two sets of hypernyms corresponding to two different senses.

Sense 1: *cantaloupe→sweet melon→melon→gourd →plant→ organism→Living thing*
Sense 2: *cantaloupe→sweet melon→melon→edible fruit→green goods→ food*

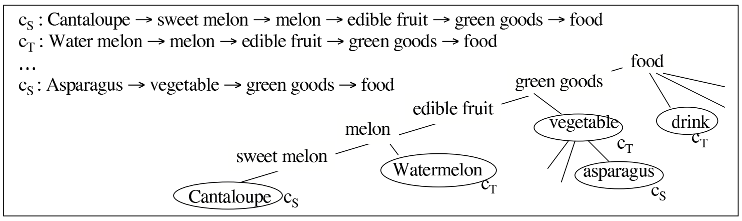
c_S : Cantaloupe → sweet melon → melon → edible fruit → green goods → food
c_T : Water melon → melon → edible fruit → green goods → food
...
c_S : Asparagus → vegetable → green goods → food
food
green goods
edible fruit
melon
sweet melon
Cantaloupe
c_S
Watermelon
c_T
vegetable
c_T
asparagus
c_S
drink
c_T

Fig. 3. A sub-graph of T_{wn} where *cantaloupe* and *watermelon* are related

Only the paths from the invoked terms to $root_A$ will be selected because they represent the only senses which are accurate for the application (sense 2 in the example, the application root node being *food*). So a sub-tree, denoted T_{wn}, is obtained. It is composed of all the terms and the relations of the retained paths (cf.Fig.3). For each concept c_S, STR_W selects in T_{wn} the most similar concept belonging to T_T using $Wu\&Palmer$'s measure [5].

$$Sim_{W\&P}(c_1, c_2) = \frac{2 * depth\ (LCA(c_1, c_2))}{depth(c_1) + depth\ (c_2)}$$

According to the $sim_{W\&P}$ measure, the concept that is the most similar to a node c_S is its parent. Moreover, we showed in [3] that the similarity is higher between c_S and any of its siblings or any of the descendants close to its siblings than between c_S and its grandparent, until a depth p that can be computed for each node c_S in function of its depth in the tree. In the same way, we can compute the depth p' from which the similarity of the great-grandparent must be considered, and so on. Using these properties, we proposed an efficient strategy in [3] which does not require the computation of many similarity measurements.

4 Experiments and Discussion

Two kinds of experiments have been performed. First, experiments have been made in the setting of the e.dot project[1], on two real-world taxonomies in the field of predictive microbiology. Second, we applied our techniques on test taxonomies [6]. The latter are not structurally dissymmetric and cover a large domain. The application conditions of the techniques are not achieved but our objective is to make these tests in order to sketch some ideas to do improvements and to widen the scope of our approach. These experiments have shown where our specific strengths and weaknesses are. Whatever taxonomy we aligned, our approach was able to retrieve almost all the equivalence mappings given with the taxonomies. Furthermore, its strong point is to propose as a bonus a lot of other mappings (subclass mappings). Some mappings have a high precision and are then certain (likely mappings generated by the terminological techniques). Other ones (potential mappings generated by the structural techniques) are less certain (low precision) and have to be validated. This confirms the order in the application of our techniques. Concerning the structural techniques, STR_T proved to be very useful and leads to relevant mappings when concepts have labels composed of a lot of words and when some words are common to many labels. On the opposite, STR_W is all the more appropriate since the application domain is small. The real-world taxonomies which have motivated our approach gather all these characteristics, unlike the others. Better results are then obtained.

5 Conclusion

We described two structural techniques to align structurally asymmetric taxonomies. These techniques are original because different from a search of structural similarity in models. They are executed to suggest additional mappings. These mappings are not certain but they can be a good complement, if human involvement is possible, as experiments showed. We will continue this work by adapting and extending our techniques according to the experiment results. Our first objective is to be able to align taxonomies relative to larger application domains.

References

1. D. Lin. An Information-Theoretic Definition of Similarity, In *Proc. of the Int. Conf. on Machine Learning (ICML)*, pp. 296-304, 1998.
2. Miller, G. A. WordNet: A lexical Database for English. Communications of the ACM. (1995) Vol. 38(11) 39-45
3. C. Reynaud, B. Safar. Structural Techniques for Alignment of Taxonomies: experiments and evaluation, In *TR 1453, LRI, Univ. of Paris-Sud*, June 2006.
4. P. Shvaiko, J. Euzenat. A Survey of Schema-based Matching Approaches, In *Journal on Data Semantics*, 2005.
5. Z. Wu and M. Palmer. Verb Semantics and Lexical Selection, In *Proc. of 32nd Meeting of the Ass. for Computational Linguistics*, 1994.
6. http://www.ontologymatching/evaluation.html

[1] E.dot is a research project funded by national network on software technology (RNTL), 2003-2005.