TaxoMap in the OAEI 2009 alignment contest

Fayçal Hamdi[1], Brigitte Safar[1], Nobal B. Niraula[2], and Chantal Reynaud[1]

[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] nobal.niraula@inria.fr

Abstract. TaxoMap is an alignment tool which aims to discover rich correspondences between concepts. 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 reduces significantly runtime and enables parameterization by specifying the ontology language and different thresholds used to extract different mapping relations. It improves terminological techniques, with a better use of TreeTagger and introduces new structural techniques which take into account the structure of ontology. Special effort has been made to handle large-scale ontologies by partitioning input ontologies into modules to align. We conclude the paper by pointing out the necessary improvements that need to be made.

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.

To find mappings in this context, we can only use the following available elements: labels of concepts and hierarchical structures.

The new implementation of TaxoMap proposes a better morpho-syntactic analysis and new techniques. Moreover, the methods to partition large ontologies into modules which TaxoMap can handle easily were refined.

We take part to five tests. We hope we perform better in terms of precision of mappings generated and runtime. Tests on library data sets allow us to experiment our algorithm on large multilingual ontologies (English, French, and German).

2 Presentation of the System

2.1 State, Purpose and General Statement

We consider an ontology as a pair (C, H_C) consisting of a set of concepts C arranged in a subsumption hierarchy H_C. A concept c is defined by two elements: a set of labels and subclass relationships. The labels are terms that describe entities in natural language and which can be an expression composed of several words. A subclass relationship establishes links with other concepts.

Our alignment process is oriented; from a source (O_S) to a target (O_T) ontology. It aims at finding one-to-many mappings between single concepts and establishing three types of relationships, equivalence, subclass and semantically related relationships defined as follows.

Equivalence relationships An equivalence relationship, *isEq*, is a link between a concept in O_S and a concept in O_T with labels assumed to be similar.

Subclass relationships Subclass relationships are usual *isA* class links. When a concept c_S of O_S is linked to a concept c_T of O_T with such a relationship, c_T is considered as a super concept of c_S.

Semantically related relationships A semantically related relationship, *isClose*, is a link between concepts that are considered as related but without a specific typing of the relation.

2.2 Techniques Used

The different techniques are based on the use of the moropho-syntactic analysis tool TreeTagger [1], and a similarity measure which compares the trigrams of the concept labels [2].
TreeTagger is a tool for tagging text with part-of-speech and lemma information, enables to take into account the language, lemma and an use word categories in an efficient way. The words are classified as functional (verbs, adverbs or adjectives) and stop words (articles, pronouns). Once classified by TreeTagger, the words are divided into two classes, **full words** and **complementary words**, according to their category and their position in the label. In principle, all names are full words except if they are placed after a determiner, all other words are complementary words.
This classification is then used to give more weight to the full words in the calculation of similarity between labels.

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_s and one label of c_t is greater than a threshold (Equiv.threshold).

- High lexical similarity: Let c_{tmax} be the concept in O_T with the highest similarity measure with c_s. If the similarity measure is greater than a threshold (HighSim.threshold) and if one of the labels of c_{tmax} shares at least two full words in common with one of the labels of c_s, the heuristic generates the relationship $< c_s\ isA\ c_{tMax} >$ if the label of c_{tmax} is included in the c_s one, otherwise it generates $< c_s\ isClose\ c_{tMax} >$.
- 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 $< c_s\ isA\ c_{tmax} >$. Inversely, if one of the labels of c_s is included in one of the labels of c_{tmax}, we propose a semantically related relationships $< c_s\ isClose\ c_{tmax} >$.
- 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 three following techniques can be used:
  - the relationship $< c_s\ isClose\ c_{tMax} >$ is generated if one of the labels of c_s is included in one of the labels of c_{tMax}, and the words of the included label are complementary words.
  - the relationship $< c_s\ isClose\ c_{tMax} >$ 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).
- Reasoning on structure:
  - an isA relationship $< c_s\ isA\ c_t >$ is generated if the subclass relation $< c_s\ isSubClassOf\ X >$ appears in O_S and if the equivalence mapping $< X\ isEq\ c_t >$ have been identified.
  - the relationship $< c_s\ isClose\ c_t >$ is generated if c_t is the concept in O_T which have the most number of children in O_T with the same label as the children of c_s in O_S. More details of this approach are given at the end of this sub-section.
  - an isA relationship $< c_s\ isA\ p >$ is generated if the three concepts in O_T with the highest similarity measure with c_s have similarity greater than a threshold (Struct.threshold), and have a common father p in O_T.

As we mentioned above, we use a structural heuristic based on the $Semantic\ Cotopy$ measure of a concept, proposed by Maedche and Staab [3]. The $Semantic\ Cotopy$ is based on the intentional semantics of a concept c in an ontology O, $SC(c, O)$, defined as the set of all its super- and sub-concepts in O. When a concept c belongs to two ontologies, one can define the taxonomic overlap (TO) between O_1 and O_2 for this concept, denoted $TO(C, O_1, O_2)$ and defined as the ratio between the number of common elements in the intentional semantics of c in O_1 and in O_2 and the total number of elements belonging to the union of these two sets. If a concept c is in O_1 but not in O_2, an optimistic approximation of $TO(c, O_1, O_2)$ is defined as the maximum overlap obtained by comparing $SC(c, O_1)$ to the intentional semantics of all the concepts in O_2. Our heuristic uses $SC_D(c)$ which includes only the concept and its descendants instead of the original Semantic Cotopy. If a concept c is in O_1 but not in O_2, we propose as candidate mapping for this concept c, the concept c_{Max} of O_2 which maximizes the TO, if c and c_{Max} have at least two descendants in common.

2.3 Partitioning of large scale ontologies

We propose a method of ontology partitioning [4], that relies on the implementation of PBM [5] algorithm. PBM partitions large ontologies into small blocks (or modules) and constructs mappings between the blocks, using predefined matched class pairs, called *anchors* to identify related blocks. We reuse the partitioning part and the idea of anchors, but the originality of our method, called PAP (*Partition, Anchor, Partition*), is that it is *alignment oriented*, that means that the partitioning process is influenced by the mapping process.

The PAP method consists of:

- decompose the most structured ontology, that will be called the *target*, O_T, into several blocks B_{Ti}, according to the PBM algorithm.
- force the partitioning of the other ontology, called the *source* O_S, to follow the pattern of O_T. To achieve this, the method identifies for each block B_{Ti} constructed from O_T all the anchors belonging to it. Each of these sets of anchors will constitute the kernel or *center* CB_{Si} of a future block B_{Si} which will be generated from the source O_S.
- reuse the PBM algorithm to partition the source O_S around the centers CB_{Si}.
- align each block B_{Si} built from a center CB_{Si} with the corresponding block B_{Ti}.

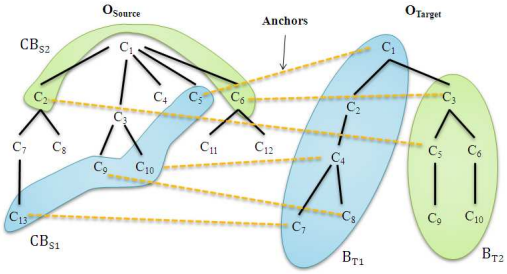

Fig. 1. *The centers CB_{Si} identified from B_{Ti}*

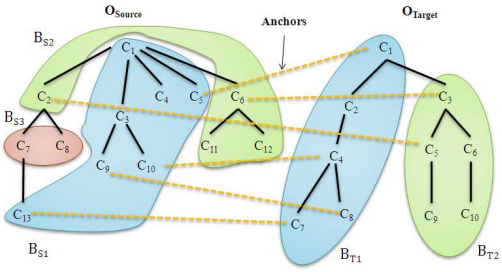

Fig. 2. Partition of O_S around the centers CB_{Si}

The tests show that the maximum size of the blocks has to be fixed for the target ontology. If the themes covered by both ontologies are of the same importance, i.e. if the source ontology corresponds to a representation of the same importance than the representation of the target one, a maximum size for the blocks in the source ontology is not needed. Their size will become close to the size of the blocks of the target ontology. This phenomenon allows to avoid obtaining a lot of small isolated blocks which appear when the maximum size of the blocks of the source ontology is fixed.

So, on the example of Fig2, the B_{S3} block remains isolated because the size of of the source blocks was fixed. Without limitation of the size, the B_{S3} block can be merged with B_{S2}. The only blocks which will remain isolated will be the blocks built

when the source ontology will be partitioned, independently of the kernels identified in the decomposition of the target ontology, i.e. concepts with no relation with those of the target ontology. So, the fact that the concepts belonging to these isolated blocks are not aligned should not damage our results.

2.4 Adaptations made for the Evaluation

Unlike in previous years, we have made some specific adaptations for the OAEI 2009 campaign.

For Anatomy task, we did not use the techniques which generate isA relationship. All the alignments outputted by TaxoMap are uniformly based on the same parameters. We had, however, fixed confidence values depending on relation types.

For library test, data sets consist of multilingual ontologies. In order to use lexical comparison, we translated non-English labels of all of the concepts of the vocabularies into English. The translation is done by using Googles translation APIs.

2.5 Link to the system and parameters file

TaxoMap requires:

- Mysql [3]
- Java (Version 1.5 and above)[4]
- Google's Java Client API for Translation [5]
- TreeTagger with its language parameter files [6]

The version of TaxoMap (with parameter files) used in 2009 contest can be downloaded from:

- http://www.lri.fr/~hamdi/TaxoMap.jar: a parameter lg has to be specified it denotes the language of the ontology. For example TaxoMap.jar *fr* to perform alignment on ontologies in French. If no language is specified, it is supposed to be English.
- http://www.lri.fr/~hamdi/TaxoMap.properties: a parameter file which specifies:
  - The command to launch TreeTagger.
  - TreeTagger word categories that has to be considered as functional, stop words and prepositions.
  - The RDF output file.
  - Different thresholds of similarity, depending on the method used.
- http://www.lri.fr/~hamdi/dbproperties.properties: a parameter file which contains the user and password to access to MySql.

[3] http://www.mysql.com
[4] http://java.sun.com
[5] http://code.google.com/p/google-api-translate-java
[6] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

2.6 Link to the Set of Provided Alignments

The alignments produced by TaxoMap are available at the following URLs:
http://www.lri.fr/~hamdi/benchmarks/
http://www.lri.fr/~hamdi/anatomy/
http://www.lri.fr/~hamdi/directory/
http://www.lri.fr/~hamdi/library/
http://www.lri.fr/~hamdi/benchmark-subs/

3 Results

3.1 Benchmark Tests

Since our algorithm only considers labels and hierarchical relations and only provides mapping for concepts, the recall would have been low even for the reference alignment. The overall results would have been similar -with no surprise- to those of last year.

3.2 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. As last year, we considered *Human* as the target ontology as is it well structured and larger than *Mouse*.

TaxoMap performs the alignment (with no need to partition) in about 8 minutes which is better than last year [6] where TaxoMap took about 25 minutes to align the two ontologies.

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. As a result, we found fewer mappings than last year but we hope that the precision will be better.

- For the first task, TaxoMap discovers 1274 mappings, 973 Equivalence relations and 301 Proximity relations.
- For the second task, we got only 1084 mappings, 973 Equivalence relations and 111 Proximity relations, using only the heuristic which identifies the relation $< c_s \ isClose \ c_{tMax} >$ when one of the labels of c_s is included in one of the labels of c_{tMax}.
- For the third task, we used, in addition of the techniques ot the first task, the heuristic which identifies subsumption links with "High Lexical Similarity". This allows to discover 1451 mappings and to slightly increase the recall, but reduce the precision. In fact, many mappings like <hand blood vessel isA Blood Vessel > or <iris blood vessel isA Blood Vessel > are semantically correct but become false when the subsumption relation isA is automatically replaced by an Equivalence relation.

– For the fourth task, we used the partial reference mapping in our partitioning method and we obtained 1131 mappings. This lower number of mapping is explained by two facts. The first one is that the structural heuristic based on the $Semantic\ Cotopy$ is the only one of which the results can be improved by the use of the partial mapping. The second one is that the partitioning method increases the precision but reduces the recall.

3.3 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.

3.4 Library Test

In order to use lexical comparison in library data sets, which consist of multilingual ontologies, we used Google translation API [7] to translate non-English labels into English. With our current configuration, we cannot partition the large sized library ontologies. However, we used just a part of its data set to partition and then to find the mappings among concepts.

As skos relations will be evaluated, we change different mapping types to skos ones with these confidence values:

– (type1) *isEq* relations become skos:exactMatch with a confidence value set to 1.
– (type2) *isA* relations become skos:narrowMatch with a confidence value set to 1 for label inclusion, 0.5 for relations generated by structural technique or by relative similarity method.
– (type3) *isGeneral* relations become skos:broadMatch with a confidence value set to 1.
– (type4) *isClose* relations become skos:relatedMatch with a confidence value set to 1.

Generated mappings are as follows:

– ***LCSH-RAMEAU:*** 5074 *type1* relations, 48817 *type2* relations, 116789 *type3* relations and 13205 *type4* relations.
– ***RAMEAU-SWD:*** 1265 *type1* relations, 6690 *type2* relations, 17220 *type3* relations and 1317 *type4* relations.
– ***LCSH-SWD:*** 38 *type1* relations.

3.5 Benchmark-Subs Test

Benchmark-Subs tests aims to evaluate alignments which contain other mapping relations than equivalence. Two tasks are available in this test: Gold-standard based evaluation concerning the evaluation of subsumption relations and open-ended task concerning the evaluation of equivalence and non-equivalence mappings. In our tool, for the first task, we use lexical methods to obtain subsumption relations.

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. Our partitioning algorithms allow us to participate to tests with large ontologies.

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 hierarchicals ones.
- Use of 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 [8] [9].

5 Conclusion

This paper reports our participation to OAEI campaign with the new implementation of TaxoMap. Our algorithm proposes an oriented mapping between concepts. Due to partitioning, it is able to perform alignment on real-world ontologies. Our participation in the campaign allows us to test the robustness of TaxoMap, our partitioning algorithms 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] Hamdi, F., Safar, B., Reynaud, C. and Zargayouna, H. Alignment-based Partitioning of Large-scale Ontologies, in Advances in Knowledge Discovery and Management (AKDM09), to appear.

[5] Hu, W., Zhao, Y., and Qu, Y. Partition-based block matching of large class hierarchies, Proc. of the 1st Asian Semantic Web Conference (ASWC06). pp.72-83, (2006)

[6] Hamdi, F., Zargayouna, H., Safar, B., and Reynaud, C. TaxoMap in the OAEI 2008 alignment contest, Proceedings of the ISWC'08 Workshop on Ontology Matching OM-08 (2008)

[7] http://code.google.com/p/google-api-translate-java/

[8] Reynaud, C. and Safar, B. When usual structural alignment techniques don't apply, The ISWC'06 Workshop on Ontology matching (OM-06), (2006)

[9] Reynaud, C. and Safar, B. Exploiting WordNet as Background Knowledge,The ISWC'07 Workshop on Ontology Matching (OM-07), (2007)