TaxoMap in the OAEI 2008 alignment contest

Fayçal Hamdi[1], Haïfa Zargayouna[2], Brigitte Safar[1], and Chantal Reynaud[1]

[1] LRI, Universit Paris-Sud, Bt. G, INRIA Futurs
2-4 rue Jacques Monod, F-91893 Orsay, France
firstname.lastname@lri.fr
[2] LIPN, Université Paris 13 - CNRS UMR 7030,
99 av. J.B. Clément, 93440 Villetaneuse, France.
haifa.zargayouna@lipn.univ-paris13.fr

Abstract. TaxoMap is an alignment tool which aim is 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. Our participation in last year edition of the competition have put the emphasis on certain limits. TaxoMap 2 is a new implementation of TaxoMap that reduces significantly runtime and enables parameterization by specifying the ontology language and different thresholds used to extract different mapping relations. The new implementation stresses on terminological techniques, it takes into account synonymy, and multi-label description of concepts. Special effort was 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 ressources (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.

Previous participation of TaxoMap in the alignment contest [2], despite positive outcome, have put the emphasis on certain limits:

- Multi-label concepts: previous version of TaxoMap assumed that a concept has only one label. This leads to loose interesting relations between multi-label concepts.
- Large ontologies: TaxoMap were unable to run on real ontologies, such as Agrovoc[3].

[3] http://www4.fao.org/agrovoc/

TaxoMap 2 is a new implementation of TaxoMap which aims to overcome these limits and provides modular code (easily extensible). It introduces a morphosyntactic analysis and new heuristics. Moreover, we propose new methods to partition large ontologies into modules which TaxoMap can handles easily.

We take part to four tests. Results on benchmarks are almost the same as last year as the philosophy behind TaxoMap reminds the same (oriented alignment, between concepts only). We perform better -in terms of number of mappings generated and runtime- on Anatomy. Library test allows us to perform a new algorithm for ontology partitioning and to experiments our system with a new language (Dutch). Directory test enables to test our alignment tool in real world taxonomy integration scenario.

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_{Source}) to a target (O_{Target}) 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_{Source} and a concept in O_{Target} with labels assumed to be similar.

Subclass relationships Subclass relationships are usual *isA* class links. When a concept c_S of O_{Source} is linked to a concept c_T of O_{Target} 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

TaxoMap 2 improves terminology alignment techniques. The use of TreeTagger [3], 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. TaxoMap performs a linguistic similarity measure between labels of concepts. The measure takes into consideration categories of words which compose a label. The words are classified as functional (verbs, adverbs or adjectives) and stop words (articles, pronouns).

Stop words category enables to ignore these words in similarity computation. Functional words has less power than all the other (noun, etc.). The position of a word in the label is also of importance, a common word between two labels is less important after a preposition than a word that is a head. TreeTagger, however, is error-prone, due essentially to short labels.

The main methods used to extract mappings between a concept c_s in O_{Source} and a concept c_t in O_{Target} 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).
- Label inclusion (and its inverse) and hidden inclusion: Then, we consider inclusion of label words: let c_t be the concept in O_{Target} with the highest similarity measure with c_s. If one of the labels of c_t is included in one of the labels of c_s, we propose a subclass relationship $< c_s\ isA\ c_t >$. Inversely, if one of the labels of c_s is included in one of the labels of c_t, we propose a semantically related relationships $< c_s\ isGeneral\ c_t >$. If c_t is not the concept with the highest similarity measure, its measure must be greater than a threshold (HiddenInc.thresholdSim) and the highest similarity measure must be greater than another threshold (HiddenInc.thresholdMax). The intuition behind this strategy is to extract hidden inclusion.
- Reasoning on similarity values : Let c_{tMax} and c_{t2} be the two concepts in O_{Target} with the highest similarity measure with c_s, the relative similarity is the ratio of c_{t2} similarity on c_{tMax} similarity. 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 the similarity of c_{tMax} is greater than a threshold (isCloseBefore.thresholdMax) and if one of the labels of c_s is included in one of the labels of c_{tMax}.
  - 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 is generated if the three concepts in O_{Target} with the highest similarity measure with c_s have similarity greater than a threshold (Struct.threshold), and has a common father.

2.3 Partitioning of large scale ontologies

We propose two methods of ontology partitioning. The aim of our methods is to have minimum blocs to align with maximal number of concepts (that TaxoMap is able to handle). The originality of our methods is that they are *alignment oriented*, that means that the partitioning process is influenced by the mapping process.

The two methods relies on the implementation of PBM[4] algorithm. PBM partitions large ontologies into small blocks (or modules) and construct mappings between the blocks, using predefined matched class pairs, called *anchors* to identify related blocks. We only reuse the partitioning part and the idea of anchors, but adapt them in order to take into account the alignment process in the partitionning. We identify the set of *anchors* as the set of concepts which have the same label in the two ontologies. Even on very large ontologies, this set is computable with a fast and strict equality measure. We also used the possible dissymmetry between ontologies to order the partitionning: if one ontology is well-structured, it will be easier to split it up into cohesive modules, and its partitionning can be used as guideline to partition the other ontology.

The methods proposed are as follows:

- Method1 (see figure 1):
  1. Use PBM algorithm to partition the target ontology O_T into some blocs B_{Ti}.
  2. Identify the set of anchors included in each module B_{Ti}. This set will be the kernel or *center* CB_{Si} of the future module B_{Si} which will be generated from the source ontology O_S.
  3. Use PBM algorithm to partition the source ontology around the identified centers CB_{Si}.
  4. Align each module B_{Si} with the corresponding module B_{Ti}.
- Method2 (see figure 2):
  1. Partition the target ontology O_T by modifying PBM algorithm in order to take into account anchors. Generated modules contain coherent set of concepts that maximize the number of anchors.
  2. Partition the source ontology O_S the same way then step 1. The interesting anchors that influence partitioning are those that goes in the same module generated from O_T.
  3. Align modules that share maximal number of anchors.

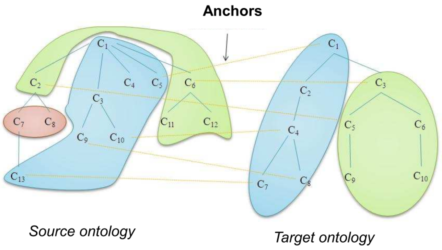

Fig. 1. *Method1 for partitioning*

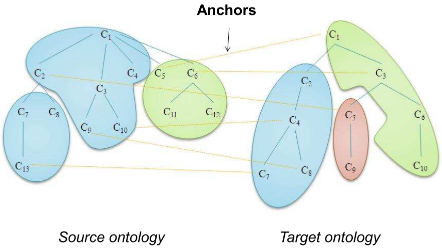

Fig. 2. *Method2 for partitioning*

2.4 Adaptations made for the Evaluation

We do not make any specific adaptation in the OAEI 2008 campaign. All the alignments outputted by TaxoMap are uniformly based on the same parameters. For library test, the language was set to *nl* (for Dutch). We had, however, fixed confidence values depending on relation types.

2.5 Link to the system and parameters file

TaxoMap requires :

- Mysql
- Java (from 1.5)
- TreeTagger [4] with its language parameter files.

The version of TaxoMap (with parameter files) used in 2008 contest can be downloaded from:

- http://www.lri.fr/˜haifa/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/˜haifa/TaxoMap.properties: a parameter file which specifies:
  - The command to launch tree-tagger.
  - 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/˜haifa/dbproperties.properties: a parameter file which contains the user and password to access to MySql.

2.6 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/
http://www.lri.fr/˜haifa/directory/
http://www.lri.fr/˜haifa/library/

3 Results

3.1 Benchmark Tests

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

The whole process of alignment costs less than 2 minutes.

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

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 3,044 classes. As last year, we considered *Human* as the target ontology as is it well structured and larger than *Mouse*.
TaxoMap gains considerably on runtime, it performs the alignment (with no need to partition) in about 25 minutes which is better than last year where TaxoMap took about 5 hours to align the two ontologies.

TaxoMap generates much more mappings than last year. Only about 200 concepts were left unmapped, whereas last year it was nearly 900.
As only equivalence relationships will be evaluated, we change different mapping types to equivalence with these confidence values:

- (type1) For *isEq* and *isClose* relations, confidence value was set to 1.
- (type2) For *isA* relations generated by label inclusion, confidence value was set to 0.8.
- (type3) For *isA* relations generated by structural technique or by relative similarity method, confidence value was set to 0.5.

TaxoMap discovers 2 533 mappings: 1 208 type1 relations, 1 190 type2 relations and 135 type3 relations. The improvement in comparaison with last year results relies on the use of TreeTagger and on taking into account synonymy.

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

The library task includes two SKOS thesauri GTT and Brinkman thesauri. Since TaxoMap focuses on Web ontologies expressed in RDFS and OWL, we have to adopt two OWL version ontologies transformed by campaign organizers in this task. GTT owns 35,000 classes, while Brinkman thesauri owns 5,000 classes. The main drawback of using OWL ontologies is that there is no distinction in OWL descriptions (rdfs:label statements) between skos:prefLabel, skos:altLabel and skos:hiddenLabel statements, which removes the subtle distinctions that exist between these different properties.

We applied the first method of partitioning, this is due to the fact that only 3535 anchors were discovered and that the two ontologies were poorly structured. As the method2 relies on these two informations simultaneously, the partitioning results were not judged relevant.

The partitioning of Brinkman thesauri (considered as target ontology) leads to 227 modules, the largest module contains 703 concepts. GTT (source ontology) is partitionned into 18 306 modules, 16 265 modules contain only one concept, the largest module contains 517 concepts. We performed 212 combinations that leads to 3 217 mappings.

The fact that the total number of mappings is less than the number of found anchors is due to the fact that anchors are computed between labels (a concept described by three labels can have three anchors, which is not the case for mappings, where a concept is matched to only one concept). As alignments are performed between modules, this can lead to loose some potential mappings. This is particularly the case of all modules that contain only one concept, as they are ignored by the alignment process.

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: 1 872 type1 relations, 1 031 type2 relations, 274 type3 relations and 40 type4 relations. The whole process of alignment costs about 40 minutes. The partitioning process costs nearly 2 hours.

The language of both thesauri is Dutch, we launched tree-tagger with Dutch parameter file. The main difficulty is that there were no Tagset description given for this language and it was difficult to specify word categories needed for the linguistic similarity method.

4 General Comments

4.1 Results

TaxoMap 2 significantly improves the results on the previous version of TaxoMap in terms of runtime and number of generated mappings. The new implementation offers extensibility and modularity of code. TaxoMap can be parameterized by the language used in ontologies and different thresholds. We put the emphasis on terminological alignment by taking into account synonymy and multi-label concepts. Our partitioning algorithms allows us to participate to tests with large ontologies.

4.2 Future Improvements

The following improvements can be made to obtain better results:

- 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 [5] [6].

5 Conclusion

This paper reports our participation to OAEI campaign with a new implementation of TaxoMap. Our algorithm proposes an oriented mapping between concepts. TaxoMap 2 is better now than last year. 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 terminological heuristics.

References

[1] Lin, D. : An Information-Theoretic Definition of Similarity. ICML. Madison. (1998) 296–304

[2] Zargayouna, H., Safar, B., and Reynaud, C. : TaxoMap in the OAEI 2007 alignment contest Proceedings of the ISWC'07 workshop on Ontology Matching OM-07 (2007) 268-275

[3] Schmid H. : Probabilistic Part-of-Speech Tagging Using Decision Trees. International Conference on New Methods in Language Processing (1994)

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

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

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