AROMA results for OAEI 2009

Jérôme David[1]

Université Pierre-Mendès-France, Grenoble
Laboratoire d'Informatique de Grenoble
INRIA Rhône-Alpes, Montbonnot Saint-Martin,
France
Jerome.David-at-inrialpes.fr

Abstract. This paper presents the results obtained by AROMA for its second participation to OAEI. AROMA is an hybrid, extensional and asymmetric ontology alignment method that makes use of the association paradigm and a statistical interestingness measure, the implication intensity. AROMA performs a post-processing step that includes a terminological matcher. This year we modify this matcher in order to improve the recall obtained on real-case ontology, i.e. anatomy and 3xx tests.

1 Presentation of AROMA

1.1 State, purpose, general statement

AROMA is an hybrid, extensional and asymmetric matching approach designed to find out relations of equivalence and subsumption between entities, i.e. classes and properties, issued from two textual taxonomies (web directories or OWL ontologies). Our approach makes use of the association rule paradigm [Agrawal *et al.*, 1993], and a statistical interestingness measure. AROMA relies on the following assumption: *An entity A will be more specific than or equivalent to an entity B if the vocabulary (i.e. terms and also data) used to describe A, its descendants, and its instances tends to be included in that of B.*

1.2 Specific techniques used

AROMA is divided into three successive main stages: (1) The pre processing stage represents each entity, i.e. classes and properties, by a set of terms, (2) the second stage consists of the discovery of association rules between entities, and finally (3) the post processing stage aims at cleaning and enhancing the resulting alignment.

The first stage constructs a set of relevant terms and/or datavalues for each class and property. To do this, we extract the vocabulary of class and property from their annotations and individual values with the help of single and binary term extractor applied to stemmed text. In order to keep a morphism between the partial orders of class and property subsumption hierarchies in one hand and the inclusion of sets of term in the other hand, the terms associated with a class or a property are also associated with its ancestors.

The second stage of AROMA discovers the subsumption relations by using the association rule model and the implication intensity measure [Gras *et al.*, 2008]. In the context of AROMA, an association rule $a \rightarrow b$ represents a quasi-implication (i.e. an implication allowing some counter-examples) from the vocabulary of entity a into the vocabulary of the entity b. Such a rule could be interpreted as a subsumption relation from the antecedent entity toward the consequent one. For example, the binary rule $car \rightarrow vehicle$ means: "The concept car is more specific than the concept $vehicle$". The rule extraction algorithm takes advantage of the partial order structure provided by the subsumption relation, and a property of the implication intensity for pruning the search space.

The last stage concerns the post processing of the association rules set. It performs the following tasks:

- deduction of equivalence relations,
- suppression of cycles in the alignment graph,
- suppression of redundant correspondences,
- selection of the best correspondence for each entity (the alignment is an injective function),
- the enhancement of the alignment by using a string similarity -based matcher and previously discovered correspondences.

This year, we made some changes on the string similarity -based matcher. These changes are primarily designed to improve the recall on anatomy track. Now AROMA includes an equality -based matcher: two entities are considered equivalent if they share at least one annotation. This matcher is only applied on unaligned pairs of entities.

The string similarity based matcher still makes use of Jaro-Winkler similarity but relies on a different weighting scheme. As an ontology entity is associated to a set of annotations, i.e. local name, labels and comments, we need a collection measure for aggregating the similarity values between all entity pairs. Last year, we relied on maximal weight maximal graph matching collection measure, see [David and Euzenat, 2008] for details.

In order to favour the measure values of most similar annotations pairs, we choose to use the following collection measure:

$$\Delta_{mw}(e, e') = \begin{cases} \frac{\sum_{a \in T(e)} \arg\max_{a' \in T(e')} sim_{jw}(a,a')^2}{\sum_{a \in T(e)} \arg\max_{a' \in T(e')} sim_{jw}(a,a')} & \text{if } |T(e)| \leq |T(e')| \\ \Delta_{mw}(e', e) & \text{otherwise} \end{cases}$$

where $T(e)$ is the set which contains the annotations and the local name of e, and sim_{jw} is the Jaro-Winkler similarity. For all OAEI tracks, we choose a threshold value of 0.8.

For more details about AROMA, the reader should refer to [David *et al.*, 2007; David, 2007].

1.3 Link to the system and parameters file

The version 1.1 of AROMA has been used for OAEI2009. This version can be downloaded at : `http://gforge.inria.fr/frs/download.php/23649/AROMA-1.1.zip`.

The command line used for aligning two ontologies is:

```
java -jar aroma.jar onto1.owl onto2.owl [alignfile.rdf]
```

The resulting alignment is provided in the alignment format.

1.4 Link to the set of provided alignments (in align format)

```
http://www.inrialpes.fr/exmo/people/jdavid/oaei2009/results_AROMA_oaei2009.zip
```

2 Results

We participated to the benchmark, anatomy and conference tracks. We used the same configuration of AROMA for all tracks. We did not experience scaling problem. Since AROMA relies on syntactical data without using any multilingual resources, it is not able to find alignment on the multilingual library track. Finally, we also did not participate either to the instance matching track since AROMA is not designed for such a task.

2.1 Benchmark

Since AROMA mainly relies on textual information, it obtains bad recall values when the alterations affect all text annotations both in the class/property descriptions and in their individual/property values. AROMA does not seem to be influenced by structural alterations (222-247). On these tests, AROMA favours high precision values in comparison to recall values.

In comparison with last year, the modification made on AROMA have a limited negative impact on 2xx tests. By contrast, the results on 3xx tests have been enhanced: from 82% of precision and 71% of recall to respectively 85% and 78%.

2.2 Anatomy

On anatomy test, we do not use any particular knowledge about biomedical domain. AROMA runs quite fast since it takes benefits of the subsumption relation for pruning the search space. We further optimized the code since last year and now AROMA needs around 1 min. to compute the alignment. This pruning feature used by AROMA partially explained the low recall values obtained last year. For this edition, we enhanced the recall by using also an string equality based matcher before using the lexical similarity based matcher. Since AROMA returns not only equivalence correspondences but also subsumption correspondences, its precision value is negatively influenced. It could be interesting to evaluate results by using semantic precision and recall.

3 General comments

3.1 Comments on the OAEI test cases

In this section, we give some comments on the directory and oriented matching tracks of OAEI.

Directory The two large directories, that were given in previous editions of OAEI, are divided into very small sub directories. AROMA cannot align such very small directories because our method is based on a statistical measure and then it needs some large amount of textual data. However, AROMA discovers correspondences when it is applied to the complete directories. It would be interesting to reintroduce these large taxonomies for the next editions.

Oriented matching We did not participate to this track because we think that it is not well designed. Indeed, the proposed reference alignments are not complete.

For example in the 303 test, the reference alignment contains:

- 101#MastersThesis $\leq$ 103#Academic
- 103#MastersThesis $\leq$ 101#Academic

Obviously, no reliable matching algorithm would return these two correspondences but rather:

- 101#MastersThesis $\equiv$103#MastersThesis
- 101#Academic $\equiv$ 103#Academic

In addition, from these two last correspondences, we could easily deduce the two first ones.

Our suggestion for designing a better oriented matching track would be to remove some classes and properties in the target ontologies so as to obtain complete reference alignments with some subsumption relations. For example, it would be more accurate to remove the concept MasterThesis from the ontology 103 in order to naturally change 101#MastersThesis $\equiv$103#MastersThesis by 101#MastersThesis $\leq$ 103#Academic in the reference alignment.

4 Conclusion

The version of AROMA includes a new matcher based on annotation equality. This change allows better time efficiency because it reduces the number of unaligned entities before the use of a more time consuming terminological matcher. Furthermore, we obtained better results on the 3xx tests of benchmark and tend to enhance the recall obtained on anatomy track.

References

[Agrawal *et al.*, 1993] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216. ACM Press, 1993.

[David and Euzenat, 2008] Jérôme David and Jérôme Euzenat. Comparison between ontology distances (preliminary results). In *Proceedings of the 7th International Semantic Web Conference*, volume 5318 of *Lecture Notes in Computer Science*, pages 245–260. Springer, 2008.

[David *et al.*, 2007] Jérôme David, Fabrice Guillet, and Henri Briand. Association rule ontology matching approach. *International Journal on Semantic Web and Information Systems*, 3(2):27–49, 2007.

[David, 2007] Jérôme David. *AROMA : une méthode pour la découverte d'alignements orientés entre ontologies à partir de règles d'association*. PhD thesis, Université de Nantes, 2007.

[Gras *et al.*, 2008] Régis Gras, Einoshin Suzuki, Fabrice Guillet, and Filippo Spagnolo, editors. *Statistical Implicative Analysis, Theory and Applications*, volume 127 of *Studies in Computational Intelligence*. Springer, 2008.