Comparison of textual renderings of ontologies for improving their alignment

Jorge Martinez-Gil, Ismael Navas-Delgado, Antonio Polo-Marquez and, Jose F. Aldana-Montes
University of Malaga, Department of Language and Computing Sciences, Khaos Group
Boulevard Louis Pasteur s/n, 29071 Malaga (Spain)
jfam@lcc.uma.es

Abstract

This work is about an experiment in which we have compared the textual rendering of ontologies in order to get more accurate alignments between them. The experiments we have performed consist on three main steps: rendering in a textual way two ontologies, comparing the obtained text with several algorithms for text comparing and, using the obtained result as a factor to improve the alignments between them. As result, we got some evidences that this technique gives us a good measure of the similarity of ontologies and, therefore can allow us to improve the effectiveness of the alignment process.

1. Introduction

The problem of aligning ontologies consists of finding the semantic correspondences between entities belonging to two ontologies. In the case of more than two ontologies, the problem is called multialignment, but it is not our case. More formally, the process of aligning ontologies can be expressed as a function f where given a pair of ontologies o and o', an input alignment A, a set of parameters p and a set of resources r, returns an alignment A' [1]:

$$A' = f(o, o', A, p, r)$$

Where A' is a set of mappings. A mapping is an expression that can be written in the form (e, e', n, R). Where e and e' are entities belonging to different ontologies, R is the relation of correspondence and n is a real number between 0 and 1 that represents the mathematical probability that R may be true. The entities than can be related are the concepts, roles, rules and, even axioms of the ontologies.

We wish to solve this problem in an accurate and automatic way, because it is a key aspect for getting semantic interoperability on the Semantic Web. It means that people (or groups of people) can use their own ontology without having to stick to a specific standard. It also allows them to combine their ontologies with ontologies of partners in an easy and secure way.

The reminder of this article is as follow: Next section describes briefly the state of the art on ontology alignment, from the point of view of the techniques and from the point of view of the tools. Third section describes the key ideas of a new proposal and a design of an experiment to validate it. Results section shows the empirical data that we have obtained from the experiment. Discussion deals with the interpretation and application of these results. And finally, Conclusions and Future work contains the strengths and weakness of our proposal and the future improvements that are necessary to consolidate it.

2. State-of-the-art

Related to the state-of-the-art in ontology alignment, most of authors prefer explain it in two different ways: From the point of view of the techniques and from the point of view of the tools. Related to techniques and according to [2], the equivalence between entities can be seen from three main groups: *a)* based on syntactic techniques, *b)* based on semantic techniques and, *c)* based on the structure of the ontology.

Some of the most popular syntactic techniques are string metrics, string normalization and/or translation, synonyms detection and use of external resources (lexicons, thesaurus and, so on).

Related to semantics, only a few techniques have been developed. Most of them based on deductives methods. Besides, "once deductive techniques have been applied, their results might be considered as an input to inductive techniques" [2].

On structural techniques, it is important to highlight graph-based, model-based and taxonomy-based techniques, repositories of structures and statistical methods.

In this way, there are a lot of works trying to solve the problem of alignment from the three points of view and, even trying to combine them in a hybrid technique. Most of them are implemented in the form of tool, although an

0-7695-3109-1/08 $25.00 © 2008 IEEE
DOI 10.1109/CISIS.2008.71

exhaustive overview of each one of these tools overcome the boundaries of this work, we are going to show some of the most outstanding examples:

- COMA [3]. It is a generic tool that allows finding the correspondences between a wide range of schemas. It provides a library of algorithms, a module for combining the results and a platform to evaluate them. One of its strengths is the high quality of its role comparison algorithms. It allows learning and asking to the user too.
- Cupid [4]. It implements an comparison scheme algorithm that combines linguistic techniques and relations algorithms. Its operation mode consists on converting the input schemes into graphs and then using known graph algorithms.
- QOM [5]. Its philosophy consists in to find a balance between the quality of correspondences and the execution time of the task. Instead of comparing each concept of an ontology with each concept of the other ontology, first it throws heuristic functions that decrease the number of candidates. In this way, it can provide results in a short period of time.
- Anchor-Prompt [6] tries to find relationships between entities based on the primary relationships recognized before. If two pairs of terms from the ontologies are similar and there are paths connecting the terms, then the elements in those paths are often similar as well.
- S-MATCH [7]. It allows getting semantic correspondences (similarity, specialization, generalization, disjunction and overlapping) between entities that belongs to different ontologies. The system uses the notion of plug-in for extending the existent features.
- OLA [8] is behind the idea of balancing the weight of each component that compose an ontology. It converts definitions of distances based on all the input structures into a set of equations. The algorithm tries to find the ontolgy alignment that minimizes the overall distance.

Other outstanding systems are: Asco [9] that uses a combination of linguistic and structural techniques, Buster [10] that uses inference mechanisms, FCA-Merge [11] that applies techniques from natural language processing and formal concept analysis, Glue [12] that uses machine learning techniques, IF-map [13] that uses the mathemathical channel theory, Multikat [14] that implements an algorithm of comparison and integration of multiple conceptual graphs, Rondo [15] where high-level operators to manipulate models and mappings between models are defined. And finally, T-tree [16] that uses algorithms for analizing a kind of special taxonomies.

3. Problem statement

Definition 1. *Textual rendering of an ontology is the result of printing the information contained in that ontology.* It can be expressed more formally, let e an entity from an ontology O, and let $t(e)$ a function that prints the identifier of an entity, then a textual rendering T from an ontology O is an expression such:

$$\forall e \in O, \exists t(e) \Rightarrow T(O) = \{t(e)\}$$

Example 1. Textual rendering for Figure 1 is *A man is a person. A woman is a person*.

Now, we are going to explain why we think that textual renderings of ontologies are interesting.

Example 2. Note Figure 1 and Figure 2; they are very simple ontologies. They are very similar, too. For example, it is easy to align the concepts man and $woman$, using any algorithm for string matching. But, what is about $person$ and $human\ being$? We know that both represent the same object of the real world, but what computer algorithm can tell us that are the same? Based on string similarity techniques cannot. Based on taxonomy algorithms can increase the probability, but it is not enough. Based on WordNet algorithms can, but they are dangerous; imagine such concepts as 'plane' and 'aeroplane', they are synonyms, but only in some situations. We think that we can solve this problem and we are going to make an experiment to show it: Let's remember the textual rendering from the first ontology: *A man is a person. A woman is a person*.

On the other hand, textual rendering for the second ontology sample is: *A man is a human being. A woman is a human being*. Now, if we compare the two textual renderings using an algorithm as Loss of Information (LOI) [17], we have a 76.9 percent of similarity between them. We propose to use this result as a factor to increase the probability of the mappings in the output alignment.

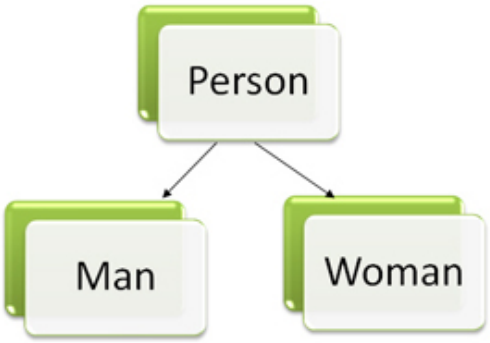

Figure 1. Ontology sample number 1

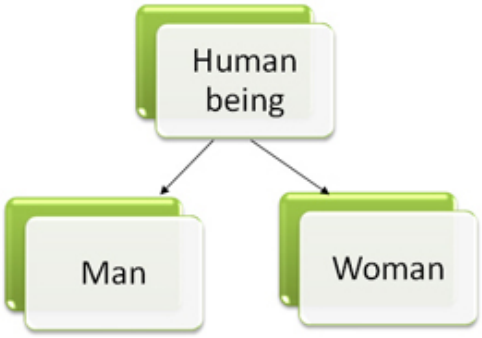

Figure 2. Ontology sample number 2

In this sense, we think that we can use this observation in order to formulate a generic technique for improving ontology mappings.

The experiment that we are going to perform consists of a previous task and then three steps. The previous task is to launch a task to align the ontologies. It is interesting to launch a simple algorithm in order (as a based on similarity string algorithm) to see how much the next steps increase the quality of the alignment. Then:

1. Rendering the ontologies.
2. Comparing the obtained text.
3. Using the result as a factor to increase the probability of the mappings may be true.

Although we have defined textual rendering already, there are several ways to render the ontology in a textual way:

Definition 2. *Crude rendering is the kind of rendering that only prints the information of the concepts and properties, excluding the relations. So it loses information about the structure.* It is good when we wish to compare only the content of the ontologies.

- **Definition 2.1**. *Partial Crude rendering is a kind of rendering used to compute the similarity rate between a concrete kind of entities in two ontologies.* It is useful in cases where concepts are very similar but other entities (properties, relations, instances, so on) are very different.
- **Definition 2.2**. *Full Crude rendering is a kind rendering used to compare the contents of the whole ontologies.* It seems to be useful when compared ontologies are very closed.

Definition 3. *Full rendering is the kind of rendering which allows to rebuild the ontology because it prints information about the content and the structure. So it is a rendering without loss of information.* It is useful in order to compare not only the contents, but the structures.

- **Definition 3.1**. *Partial Full rendering prints all the information related to a kind of entities.* As we commented earlier, it is useful when concepts are closed, but we think that there are very different instances, for example.
- **Definition 3.2**. *Complete rendering prints all the information of the ontology, so the process is reversible.*

Crude renderings try to get a measure of the resemblance of the vocabularies. In full renderings, the resemblance of vocabularies is important, but each time that a entity appear we print a more elaborated message about it. Note that the message we print is similar for the two ontologies, so we are increasing the similarity between the generated text, but also reducing the importance of the vocabularies.

In order to get empirical results from our theory, we are going to perform an experiment over two public ontologies. We have chosen the ontology about bibliography of the Institute of Information Sciences (ISI) from California, USA [18]. And the ontology about bibliography from the University of Yale [19], in the United States too. Originally, both ontologies were in DAML [20] format, but we have converted them into OWL format [21] in order to allow our software to process them. We have chosen them because we guess they have a high degree of commonality and, therefore the experiment could show us the merits of our proposal. Other important details we have considered are:

- The argument R of the mappings (relation between the entities) will be Equivalence only.
- We have determined that the degree of similarity between the textual renderings will be used for increase the n of the mappings (probability of relation between them be true).

4. Results

1. At first time, we have performed a syntactic alignment of the ontologies. We have used the Levenshtein algorithm [22]. Table 1 shows the results for the concept alignment. We have determined a low threshold for getting a significative number of pairs. Table 2 shows the results for the properties alignment. Many of them are the same in both ontologies.
2. At second time, we have performed the rendering over ontologies from the ISI and Yale. We have used Full Crude Rendering. In this way, we give more importance to the similarity of the vocabularies than to the structure of the ontologies.

ISI	Yale	n
$patent$	$Literal$	0.285
$collection$	$Incollection$	0.833
$collection$	$Publication$	0.545
$booklet$	$Incollection$	0.333
$booklet$	$Book$	0.428
$techreport$	$Techreport$	0.900
$phdthesis$	$Inproceedings$	0.307
$book$	$Book$	0.750
$manual$	$Literal$	0.285
$incollection$	$Incollection$	0.916
$incollection$	$Publication$	0.416
$conference$	$Incollection$	0.250
$proceedings$	$Inproceedings$	0.846
$inproceedings$	$Inproceedings$	0.923
$article$	$Article$	0.857
$inbook$	$Incollection$	0.250
$inbook$	$Book$	0.500

Table 1. Concept alignment. Threshold: 0.25

ISI	Yale	n
$title$	$title$	1.000
$title$	$booktitle$	0.555
$note$	$note$	1.000
$institution$	$institution$	1.000
$howpublished$	$publisher$	0.667
$editor$	$editor$	1.000
$number$	$number$	1.000
$author$	$author$	1.000
$volume$	$volume$	1.000
$location$	$Publication$	0.636
$year$	$year$	1.000
$publisher$	$publisher$	1.000
$mrnumber$	$number$	0.750
$annote$	$note$	0.666
$booktitle$	$title$	0.555
$booktitle$	$booktitle$	1.000
$edition$	$editor$	0.714
$organization$	$Publication$	0.500
$pages$	$pages$	1.000
$affiliation$	$Publication$	0.545

Table 2. Property alignment. Threshold: 0.5

3. We have used the Loss Of Information (LOI) algorithm for comparing both generated texts, we have obtained a similarity degree of 42.2 percent.

4. Finally, we have used that 42.2 percent for increase the argument n of the mappings be true (we have used the formula $n = n + (0.422 \cdot n)$. In this way, the higher values are increased significatively, while lower probabilities not. Table 3 and Table 4 shows us the new results for the concepts and the properties respectively.

In Table 5, we have extracted a statical summary from the results of our proposal[1]

As you can see, at least in this case, we have improved the precision, we have kept the recall and, of course, we have increased the F-Measure. But there are bad news too, the number of false positives has increased. We have considered that a relation is true when its n argument is equal or greater than 0.9.

Finally, we have repeated the experiment using ontologies from other fields: academic departments, people and genealogy. As you can see in Table 6, we cannot determine any kind of relation between the improved precision and the similarity of the textual renderings, but according to the performed experiments, the technique that we propose is able to improve the precision of the mappings.

[1] We have used the following formulas for the calculations:

$$Precision = \frac{Correct\ relations}{Correct\ relations + Incorrect\ relations}$$

$$Recall = \frac{Correct\ relations}{Correct\ relations + Not\ found\ relations}$$

$$F - Measure = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

5. Discussion

Note that there are a lot of concepts and properties that could be aligned using a string normalization algorithm. However, there are a few couples which couldn't. For instance: $proceedings$ and $Inproceedings$, $mrnumber$ and $number$, $collection$ and $Incollection$ and so on. Therefore, the advantages are that we have into account the similarity of the ontologies for improving the mappings. In this way, we can enrich the results generated by simple methods. We provide several ways to proceed: giving more importance to the vocabulary or giving more importance to the whole ontology. Moreover, to have into account only concrete parts of the ontologies is possible. The result of our experiment tell us that it is possible to improve the precision and F-measure of the alignment process. There are some disadvantages too; it is necessary to combine this technique with other ones, that it is to say, it is not good enough as to generate good mappings by itself. Besides, it increases the number of false positives. On other hand, you may wondered why we have not improved the recall. Think that we improve existing results, we do not look for new ones. We increase the probabilities of the relations be true, as higher are these probabilities, more be incremented and vice versa.

ISI	Yale	n (Improved)
patent	*Literal*	0.405
collection	*Incollection*	**1.000**
collection	*Publication*	0.774
booklet	*Incollection*	0.473
booklet	*Book*	0.608
techreport	*Techreport*	**1.000**
phdthesis	*Inproceedings*	0.436
book	*Book*	**1.000**
manual	*Literal*	0.405
incollection	*Incollection*	**1.000**
incollection	*Publication*	0.591
conference	*Incollection*	0.355
proceedings	*Inproceedings*	**1.000**
inproceedings	*Inproceedings*	**1.000**
article	*Article*	**1.000**
inbook	*Incollection*	0.355
inbook	*Book*	0.711

Table 3. Improved Concept alignment. Threshold: 0.25

ISI	Yale	n (Improved)
title	*title*	1.000
title	*booktitle*	0.788
note	*note*	1.000
institution	*institution*	1.000
howpublished	*publisher*	**0.946**
editor	*editor*	1.000
number	*number*	1.000
author	*author*	1.000
volume	*volume*	1.000
location	*Publication*	0.903
year	*year*	1.000
publisher	*publisher*	1.000
mrnumber	*number*	**1.000**
annote	*note*	**0.946**
booktitle	*title*	0.788
booktitle	*booktitle*	1.000
edition	*editor*	1.000
organization	*Publication*	0.710
pages	*pages*	1.000
affiliation	*Publication*	0.774

Table 4. Improved Property alignment. Threshold: 0.5

	Before	Later
Precision	63.1%	79.1%
Recall	92.3%	92.3%
$F - Measure$	74.9%	86.5%

Table 5. Summary from the experiment

Ontologies	Similarity	Precision
Departments [22] *vs* [23]	14.8%	+12.5 *p.p.*
People [24] *vs* [25]	19.2%	+8.3 *p.p.*
Bibliography [17] *vs* [18]	42.2%	+16.0 *p.p.*
Genealogy [26] *vs* [27]	61.2%	+7.6 *p.p.*

Table 6. Results obtained from alignments in other domains

But, we do not launch a alignment task again. In the experiments, we have obtained a good degree of similarity, we think that this result means that compared ontologies are similar, but we knew that we have been aligned closed ontologies. We have to study this detail more in depth in order to formulate a more accurate methodology.

6. Conclusions and future work

In this work, we have proposed a technique for getting more accurate ontology alignments. This technique is based on the comparison of the textual renderings of the ontologies to align. According to the experiments we have performed, we can conclude that comparing the textual rendering of the ontologies to align is able to improve the precision of the alignment process. However, there is work to do: At first time it is necessary to test a bigger quantity of ontologies, we are going to test the benchmark provided by the Ontology Alignment Evaluation Initiative (OAEI) [29]. Moreover, it is important to determine clearly what kind of rendering is more appropriate according to the situation, and what are the best algorithms for comparing the text obtained from the textual rendering. In this way, we wish to use not only LOI algorithm, but other text metrics.

7. Acknowledgments

This work has been funded by Spanish Ministry of Education and Science through: TIN2005-09098-C05-01.

References

[1] Jerome Euzenat, Thanh Le Bach, Jesus Barrasa, Paolo Bouquet, Jan De Bo, Rose Dieng-Kuntz, Marc Ehrig, Manfred Hauswirth, Mustafa Jarrar, Ruben Lara, Diana Maynard, Amedeo Napoli, Giorgos Stamou, Heiner Stuckenschmidt, Pavel Shvaiko, Sergio Tessaris, Sven Van Acker, and Ilya Zaihrayeu. State of the art on ontology alignment. Deliverable D2.2.3, *Knowledge web NoE*, 2004.

[2] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, 2007.

[3] H. Do and E. Rahm. *Coma - a system for œxible combination of schema matching approaches*. In Proc. VLDB, pages 610-621, 2002.

[4] J. Madhavan, P. Bernstein and E. Rahm. *Generic schema matching using Cupid*. In Proc. of the 27th VLDB Conference, pages 48-58, 2001.

[5] M. Ehrig and S. Staab. *QOM - Quick Ontology Mapping*. In Proc. 3rd ISWC, Hiroshima (JP), pages 683-697, 2004.

[6] N. Noy and M. Musen. *Anchor-prompt: using non-local context for semantic matching.* In Proc. of the workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence (IJCAI), pages 63-70, 2001.

[7] F. Giunchiglia, P. Shvaiko, and Michael Yatskevich. *S-Match: an algorithm and an implementation of semantic matching*. In Proc. of ESWS 2004, Heraklion (GR), pages 61-75, 2004.

[8] J. Euzenat and P. Valtchev. *Similarity-based ontology alignment in OWL-lite*. In Proc. 15th European Conference on Artificial Intelligence (ECAI), Valencia (SP) pages 333-337, 2004.

[9] B.T. Le, R. Dieng-Kuntz, F. Gandon. *On ontology matching problem (for building a corporate semantic web in a multi-communities organization)*. In Proc. of 6th International Conference on Enterprise Information Systems, pages 236-243, 2004.

[10] T. Vogele, S. Hubner, and G. Schuster. Buster - an information broker for the semantic web. *Knstliche Intelligenz*, **3**:31-34, July 2003.

[11] G. Stumme and A. Madche. *FCA-merge: bottom-up merging of ontologies*. In Proc. 17th IJCAI, Seattle (WA US), pages 225-230, 2001.

[12] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos and, A. Halevy. Learning to match ontologies on the Semantic Web. *The VLDB Journal*, **12**:303-319, 2003.

[13] Y. Kalfoglou and M. Schorlemmer. If-map: an ontology mapping method based on information flow theory. *Journal of data semantics*, **1**:98-127, 2003.

[14] R. Dieng and S. Hug. *Multikat, a tool for comparing knowledge from multiple experts. In Conceptual Structures: Theory, Tools and Applications*, Proc. Of the 6th Int. Conference on Conceptual Structures (ICCS'98), Montpellier (FR), pages 139-153, Springer-Verlag, 1998.

[15] S. Melnik, E. Rahm, and P. Bernstein. *Rondo: A programming platform for model management*. In Proc. ACM SIGMOD, San Diego (CA US), pages 193-204, 2003.

[16] J. Euzenat. *Brief overview of T-tree: the Tropes taxonomy building tool*. In Proc. 4th ASIS SIG/CR workshop on classification research, Columbus (OH US), pages 69-87, 1994.

[17] P. Ziegler, C. Kiefer, C. Sturm, K. R. Dittrich, A. Bernstein. *Detecting Similarities in Ontologies with the SOQA-SimPack Toolkit*. In Proc. EDBT, pages 59-76, 2006.

[18] Bibliography vocabulary from Yale University. http://www.cs.yale.edu/ dvm/daml/bib-ont.daml. Visit date: 28-dec-2007.

[19] Bibliography vocabulary from IS Institute. http://www.isi.edu/webscripter/bibtex.o.daml. Visit date: 28-dec-2007.

[20] *DARPA Agent Markup Language* (DAML). http://www.daml.org/. Visit date: 28-dec-2007.

[21] G. Antoniou and F. van Harmelen. *Web Ontology Language: OWL*. In Steffen Staab and Rudi Studer, editors, Handbook on Ontologies, pages 67-92. Springer, 2004.

[22] V. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics-Doklady*, Vol. 10, pages 707-710, August 1966.

[23] AKT Ontology. http://www.aktors.org/ontology/portal. Visit date: 28-dec-2007.

[24] Ontology for computer science academic departments. http://www.cs.umd.edu/projects/plus/DAML/onts/ cs1.0.daml. Visit date: 28-dec-2007.

[25] Ontology for describing an individual. http://daml.umbc.edu/ontologies/ittalks/person. Visit date: 28-dec-2007.

[26] Ontology that describe data from a person. http://www.cs.umd.edu/projects/plus/DAML/onts/ personal1.0.daml. Visit date: 28-dec-2007.

[27] Ontology about a subset of the GEDCOM data model. http://orlando.drc.com/daml/Ontology/Genealogy/ current/. Visit date: 28-dec-2007.

[28] Ontology about the GEDCOM data model. http://www.daml.org/2001/01/gedcom/gedcom. Visit date: 28-dec-2007.

[29] *Ontology Alignment Evaluation Initiative* (OAEI). http://oaei.ontologymatching.org/. Visit date: 28-dec-2007.