SYNTHESIS: Results for the Ontology Alignment Evaluation Initiative (OAEI) 2013

Antonis Koukourikos[1, 2], George Vouros[2], Vangelis Karkaletsis[1]

[1]Institute of Informatics & Telecommunications, NCSR "Demokritos", Greece
[2]Department of Digital Systems, University of Piraeus, Greece

Abstract. The paper presents the SYNTHESIS platform, a system for automatic ontology alignment. The system supports the model-based synthesis of different individual matching methods under a co-operational framework. The configuration that has been tested over the datasets provided by the OAEI 2013 tracks incorporates four matching methods. The paper provides a brief description of the system, presents the results acquired over the various OAEI 2013 Campaign tracks and discusses the system's strengths and weaknesses, as well as, future work that will target the observed issues.

1 Presentation of the System

1.1 State, Purpose, General Statement

Given the plethora of the different proposed approaches to ontology alignment, as well as, the variations between them in terms of the types and different facets of information they exploit, the usage (or not) of various external resources and services [1], it is evident that we need effective methods for synthesizing different matching methods.

The present paper describes a specific configuration of the SYNTHESIS platform, a system for ontology matching that combines different methods under a model-based synthesis framework [2]. The objective of SYNTHESIS is to compute coherent alignments, taking advantage of the distinct and complementary advantages of various matching methods. Subsequently we present the generic synthesis process, as well as, the individual methods integrated in the current version of the system. We proceed to present the results of SYNTHESIS over the different test sets of the OAEI 2013 campaign, and the conclusions regarding its performance, its strengths and weaknesses, and the focal points that should be taken into account for the improvement of the system.

1.2 Specific Matching Techniques Used

This section describes briefly the method for synthesizing different matching methods that is employed by SYNTHESIS. Furthermore, it describes the individual matching methods incorporated in the configuration of the system that participated in the OAEI 2013 campaign.

Synthesis. The design and initial implementation of the described matching method is described in [2]. In this work, the synthesis of different matching methods is treated as a coordination problem, aiming to maximize the welfare of the interacting entities (agents). In this setting, each agent corresponds to a specific ontology element and to an individual matching method. Each agent is responsible to decide on a correspondence for its element to a target ontology, also in coordination with the other agents, so as to preserve the semantics of specifications. An agent is characterized by: (a) its state and (b) its utility function. The state ranges in the set of those elements in the target ontology that the matching method of the agent assesses to correspond to the agent's element. A specific assignment to the state variable represents an agent's decision on a specific correspondence. Nevertheless the utility of an agent for a specific correspondence depends on the states of neighboring agents. Specifically, the utility of an agent is specified to take into account structural constraints derived from subsumption relations among classes in the source ontology. These constraints represent dependencies between agents' decisions, and must be satisfied in order for the computed correspondences to preserve the semantics of ontological specifications and ensure the coherence of the correspondences.

Actually, neighbor agents of an agent *A* are those agents that correspond to the same ontology element but to different alignment methods, as well as those agents that correspond to ontology elements that are subsumed by the ontology element of *A*.

Agents are organized in graphs where they run the max-sum algorithm [3] to compute a joined set of correspondences (i.e. an alignment) so as to maximize the sum of their utilities.

SYNTHESIS is actually a generic platform that can be configured to incorporate any number of individual matching methods.

Methods incorporated in the current version of the system. The configuration of SYNTHESIS that was used for the OAEI 2013 campaign incorporates four, most of them fairly standard, matching methods. These are described in the following subsections.

COCLU. This is a string matching technique. It is realized by a partition-based clustering algorithm, which divides the examined data (strings in our cases) into clusters and searches over the created clusters using a greedy heuristic [4]. The clusters are represented as Huffman trees, incrementally constructed as the algorithm generates and updates the clusters by processing one string at a time. The decision for adding a newly encountered string in a given cluster is based on a score function, defined as the difference of the summed length of the coded string tokens that are members of the cluster and the corresponding length of the tokens in the cluster when the examined string is added to the cluster. The implementation incorporated into SYNTHESIS exploits and compares the local names, labels and comments of the examined classes.

VSM. This is a Vector Space Models-based method [5], computing the similarity between two documents. In the case of mapping tasks, the pseudo-documents to be compared are constructed as follows: Each document corresponds to a class or property and comprises words in the vicinity of that element, i.e. all words found in (a) local name, label and comments of the class; (b) the local name, label and comments for each of the class' properties; and lexical information for its related classes, as defined in [5]. The produced documents are represented as vectors of weighted index words. Each weight is the number of words' occurrence in the document. We apply cosine similarity to measure the similarity between two vectors.

CSR. The CSR method [6] computes subsumption relationships between pairs of classes belonging in two distinct ontologies. The method treats the mapping problem as a classification task, exploiting class properties and lexical information derived from labels, comments, properties and instantiations of the compared classes. Each pair of classes is represented as a feature vector, which has a length equal to the number of distinct features of the ontologies. The classifier is trained using information of both ontologies, considering each ontology in isolation.

LDM Alignment. This new method is conceived as part of a Linked Data management system, which uses unstructured textual information from the Web, in the form of extracted relation triples, in order to perform various processes related to the whole spectrum of managing and maintaining Linked Data repositories, such as Ontology Alignment and Enrichment, Repository Population, Linkage to external repositories, and Content and Link Validation [7]. The method performs web searches, using lexical information from the local names, labels and instances of the compared classes. The web documents returned from the web searches are pre-processed in order to derive their textual information, and relation tuples are extracted from each document. The sets of relation tuples associated with each class are compared, and classes' similarity is assessed.

1.3 Adaptations Made for the Evaluation

After some preliminary runs of the system with the datasets provided by the OAEI campaign, it became evident that the main flaws of the system had to do with its inability to handle ontologies of large size (in terms of the number of elements in the ontology). This is due to the current implementation of the generic synthesis process and to the complexity of the methods incorporated in SYNTHESIS.

In order to produce a system of acceptable efficiency, we introduced a dynamic method allocation component in SYNTHESIS. The component performs a shallow analysis of the input ontologies, in terms of their size and their structure. After several runs with different method combinations for the campaign datasets, the following allocation strategy was adopted: the CSR and LDM methods were excluded when the source ontology included more than 300 classes and properties. Furthermore, CSR was excluded if the examined ontologies were relatively flat, that is if the hierarchy of classes was not deeper that three subsumption levels.

While the motivation for the introduction of this component was to obtain meaningful results for as many OAEI tracks possible, we aim to expand on the idea of dynamically invoking different sets of mapping methods, depending on the specific alignment task at hand. To this end, the method allocation component can become more intricate and analytic, and be able to select a specific configuration of mapping methods from a much larger pool, ensuring that the system has reasonable execution times while also preserving its performance in terms of precision and recall.

1.4 Link to the System and Parameters File

http://users.iit.demokritos.gr/~kukurik/SYNTHESIS.zip

1.5 Link to the set of provided alignments

http://users.iit.demokritos.gr/~kukurik/results.zip

2 Results

The subsections that follow provide an overview and a brief analysis of the results achieved by SYNTHESIS in the various tracks included in the OAEI 2013 Campaign. SYNTHESIS was packaged and executed following the setup defined by the SEALS platform and using the provided SEALS client executable JAR.

2.1 Benchmark

The following table summarizes the results obtained for the benchmark track, and specifically the bibliography test set.

Bibliographic Dataset		
Average Runtime	H-mean Precision	H-mean Recall
5217 msec	0.576	0.603

We furthermore obtained results for the finance test set, as it was provided via the SEALS platform. These results are summarized below:

Finance Dataset		
Average Runtime	H-mean Precision	H-mean Recall
974454 msec	0.504	0.605

2.2 Anatomy

SYNTHESIS was not able to finish its execution within a reasonable timeframe for this dataset.

2.3 Conference

The following table summarizes the results obtained for the conference dataset of the 2013 campaign, as they were obtained via the SEALS client. The accumulative results are as follows:

Conference Dataset		
Average Runtime	H-mean Precision	H-mean Recall
5245 msec	0.799	0.484

2.4 Multifarm

The current version of SYNTHESIS does not directly address the mapping of ontologies expressed in different languages. However, due to the fact that the synthesis approach somehow matches ontologies by respecting their hierarchical structure, the results obtained show a fairly acceptable precision. The following table summarizes the results reported for this track.

Precision	Recall	F-measure
Different Ontologies		
0.30	0.03	0.05
Same Ontologies		
0.25	0.03	0.04

2.5 Library

SYNTHESIS was not able to finish its execution within a reasonable timeframe for this dataset.

2.6 Large biomedical ontologies

SYNTHESIS was not able to finish its execution within a reasonable timeframe for this dataset.

3 General Comments

3.1 Comments on the results

As evidenced by the obtained results, the main advantages of SYNTHESIS can be summarized to the following:

- SYNTHESIS manages to balance the precision and recall throughout different datasets, even with the fairly simple matching methods running for many pairs of ontologies.

- When adequate lexical information is available, i.e. when classes' names and comments were not suppressed, SYNTHESIS is able to exploit it and produce very good results.
- The constraints taken into account by agents, enables SYNTHESIS to compute coherent alignments.

In contrast, the main drawbacks of SYNTHESIS are:

- The generic synthetic approach implemented in SYNTHESIS, does not scale well with respect to ontology size. While its runtime for small and medium size ontologies is quite satisfactory, when dealing with large or very large ontologies, the system requires a significantly bigger execution time.
- Scalability is significantly affected also by the performance of the individual matching methods incorporated in the OAEI 2013 system configuration.
- The current configuration of SYNTHESIS is sensitive to the lack of adequate lexical information for the ontology elements. In the test cases where information like local class names and labels were suppressed, the results were significantly worse. This is due to the inclusion of mainly lexical-based matching methods in the current configuration of the method.

3.2 Discussions on the ways to improve the current system

The drawbacks of the current configuration of SYNTHESIS directly lead to the main points that can be improved in the future. More specifically, the main problem in various tracks of the campaign was the fact that SYNTHESIS was not able to complete its execution within an acceptable timeframe. This motivates us to examine different scalability techniques and incorporate them in the system. The actions to improve scalability can refer to the performance of the individual methods used, as well as, the actual process of synthesizing the different methods under Synthesis.

Another important step towards improving SYNTHESIS is to design and incorporate a more intricate method for choosing individual mapping methods. This is an improvement step on itself, but it is a prerequisite for being able to introduce additional methods in Synthesis and use the ones more appropriate for a specific alignment task.

The ultimate goal is to incorporate methods that exploit different types of information available (lexical, semantic, structural) at various settings (e.g. ontologies in different languages), by performing a pre-processing step to detect the characteristics of an alignment tasks, and use the most appropriate methods for constructing the agents that will be part of the synthesis process.

4 Conclusion

The participation in the OAEI 2013 has provided significant input for the evaluation and evolution of our system. The major conclusion was the system's inability to handle ontologies of large size, which will be the focus during the immediate next steps of our research. The more detailed feedback provided by the organizers of each track was also of particular importance, as it provided further insights for the functionality and the requirements of an alignment system.

5 References

1. P. Shvaiko and J. Euzenat, "Ontology Matching: State of the Art and Future Challenges", IEEE Transactions on Knowledge and Data Engineering 2013, pp. 158-176.
2. V. Spiliopoulos and George A. Vouros, "Synthesizing Ontology Alignment Methods Using the Max-Sum Algorithm", IEEE Transactions on Knowledge and Data Engineering, vol. 24(5), pp. 940-951, May, 2012.
3. A. Farinelli, A. Rogers, A. Petcu, and N.R. Jennings, "Decentralised coordination of low-power embedded devices using the max-sum algorithm", in Proc. Of the 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), Estoril, Portugal, 2008
4. K. Kotis, A. Valarakos, and G.A. Vouros, "AUTOMS: Automating Ontology Mapping through Synthesis of Methods", in Proceedings of the OAEI (Ontology Alignment Evaluation Initiative) 2006 contest, Ontology Matching International Workshop, Athens, Georgia, USA, 2006
5. V. Spiliopoulos, A.G. Valarakos, G.A. Vouros, and V. Karkaletsis, "SEMA: Results for the ontology alignment contest OAEI 2007", OAEI (Ontology Alignment Evaluation Initiative) 2006 contest, Ontology Matching International Workshop, Busan, Korea, 2007
6. V. Spiliopoulos, G.A. Vouros, and V. Karkaletsis, "On the discovery of subsumption relations for the alignment of ontologies", Web Semantics: Science, Services and Agents on the World Wide Web, Volume 8(1), pp. 69-88, March 2010
7. A. Koukourikos, V. Karkaletsis, and G.A. Vouros, "Exploiting unstructured web information for managing linked data spaces", in Proceedings of the 17th Panhellenic Conference on Informatics (PCI '13), Thessaloniki, Greece, September 2013