ServOMBI at OAEI 2015

Nouha Kheder, Gayo Diallo

Univ. Bordeaux, ERIAS - Centre INSERM U1219, F-33000, Bordeaux, France
first.last@u-bordeaux.fr

Abstract

We describe in this paper the ServOMBI system and the results achieved during the 2015 edition of the Ontology Alignment Evaluation Initiative. ServOMBI reuse components from the ServOMap ontology matching system, which uses to participate in the OAEI campaign, and implements new features. This is the first participation of the ServOMBI in the OAEI challenge.

1 Presentation of the System

ServOMBI (ServO based Mapping with Binary Indexing) is an ontology matching system [1] which is designed by reusing the overall workflow followed by the ServOMap large scale ontology matching system [2] grounded on top of the ServO Ontology Repository (OR) system [3]. ServOMap is able to handle ontologies which contain several hundred of thousands entities. To deal with large ontologies, the system relies on terminological indexing strategy provided by the ServO OR to reduce the search space and computes an initial set of candidate mappings based on the terminological description of the entities of the input ontologies.

With ServOMBI, new components and variant algorithms have been introduced in this new version in regards to the ServOMap system. Among these new features we have : –

- a binary indexing strategy to complement the terminological indexing of ServOMap for optimizing ontology navigation,
- a modified contextual similarity computation thanks to the introduction of the binary indexing strategy during the Machine Learning (ML) step,
- a new ML algorithm during the contextual similarity
- the introduction of parallelization of some tasks for optimization purposes
- the selection of the final mappings using a variant of a stable marriage problem algorithm [8]

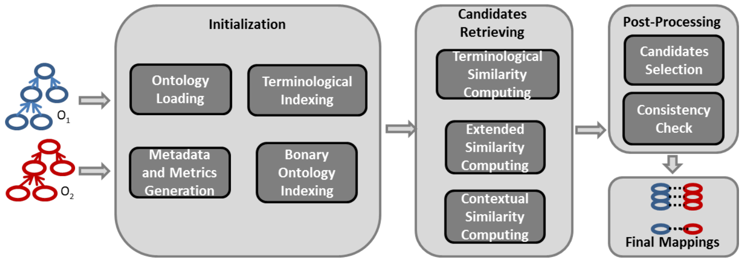

Figure 1: Overall process of ServOMBI.

In the 2015 edition, ServOMBI participated in the entity level matching tasks apart from the Multifarm task. In the following sections, we described the main characteristics of the system and the overall results obtained during this year edition of OAEI.

1.1 State, Purpose, General Statement

ServOMBI has been built on the basis of the ServOMap system. ServOMap is designed with the purpose of facilitating interoperability between different systems which are based on heterogeneous knowledge organization systems (KOS). The heterogeneity of these KOS may have several causes ranging from the language format they use to the level of formalism of the terminology which describe the entities they involve. Our system relies on Information Retrieval (IR) techniques [4] and a dynamic description of entities of different KOS for computing the similarity between them.

ServOMBI implements new features and strategies and reuse some components of the ServOMap system.

1.2 Specific techniques used

The overall process followed by the ServOMBI system is depicted on Figure 1. The initialization phase is modified by introducing the Binary Ontology Indexing (BOI).

1.2.1 Initialization phase

1. **Ontology Loading:** ServOMBI following the ServOMap approach relies on IR techniques for ontology matching. Each ontology to process is seen as a *corpus of semantic documents* to process. Each entity of the ontology is a document in the sense of IR. It is therefore necessary to identify the useful descriptors for indexing entities. The loading step perform the task of generating documents from entities. ServOMBI uses different reasoners (ELK [9], Hermit[11]) according to the size of the ontology to process.

2. **Metadata and Metrics Generation:** This step reuse the component implement in the ServOMap system and and identify 4 categories of matching tasks that are used to classify the input ontologies that are being processed : entity matching task (small, medium, big) and instances matching task.

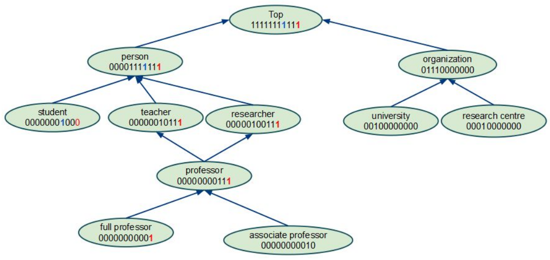

Figure 2: Example of binary indexing.

3. **Terminological Indexing:** Following a generic metamodel provided by the ServO OR, a terminological based inverted index is built from the documents generated during the loading step. ServOMBI introduces a tasks parallelization using multithreading as the input ontologies terminology indexing could be done separately.

4. **Binary Ontology Indexing:** To complement the terminological indexing and to optimize the performance in terms of processing times due to the contextual similarity computation (surrounding concepts lookups), we have introduced a binary indexing strategy for the input ontologies. This technique of taxonomical representation consists of representing each concept of the ontology by a binary code, is inspired by the CEDAR system [5]. A binary code is a number of n bits, with n the number of concepts within the processed ontology. Thus, each concept has a code (a bit vector) carrying a "1" in the position corresponding to his index and the index of any other elements that it subsumes. These bit vectors must be encoded as the reflexive transitive closure of the "is-a" relation obtained from subsort declarations. Concepts are represented by a graph. Figure 2 gives an example of a binary representation of the extract of an ontology in the academic domain. The concept *Professor* is the ancestor of *Full Professor* and *Associate Professor*. Therefore, if *Full Professor* is coded as the binary code of 1 and *Associate Professor* as the binary code of 2, *Full Professor* is coded as the binary code of 3.

1.2.2 Candidate Retrieving phase

Three main steps are used during the candidate mappings retrieving phase: terminological, extended (general purpose knowledge background) and contextual based candidate retrieving. The terminological based candidate retrieving uses indexes previously built and the IR common vectorial model. The extended candidate retrieving uses WordNet [7] while the contextual based candidate retrieving exploits the structure of each input Ontology, and the set of candidates provided by the terminological based candidate retrieving, in a ML strategy for acquiring more candidates. The ML strategy is based on the Logistic Model Trees (LMT) [10] algorithm.

Task	Precision	Recall	F-Measure
Anatomy	0.963	0.617	0.752

Table 1: Results of ServOMBI for the Anatomy track.

1.2.3 Post-Processing phase

In this phase two main steps are performed: the selction of the final mappings and consistency checking. The selection of final mappings implement an algorithm of the stable marriage problem [8].

1.3 Adaptation made for the evaluation

ServOMBI use the Lucene Apache IR library. Lucene provides functionalities for indexing and searching textual documents. The actual version of the matching system is based on the version 4 while the uploaded version for OAEI is based on the version 3.6. Their index format is slightly different. We have implemented the initial interactive matching [12] in ServOMBI using an oracle by modifying the validation process of the candidate mappings . This is performed after each round of candidate retrieving.

1.4 Link to the system and parameters file

The wrapped SEALS client for ServOMBI version used for the OAEI 2015 edition is available at http://lesim.isped.u-bordeaux2.fr/servo/ServOMBI. The instructions for testing the tool is described in the tutorial dedicated to the SEALS client[1].

1.5 Link to the set of provided alignments

The results obtained by ServOMap during OAEI 2015 are available at http://lesim.isped.u-bordeaux2.fr/servo/ServOMBI/oaei2015.zip/.

2 Results

We summarize in this section the results obtained by ServOMBI during the 2015 edition of OAEI.

2.1 Anatomy

The Anatomy track consists of finding an alignment between the Adult Mouse Anatomy and a part of the NCI Thesaurus (describing the human anatomy). The results achieved by ServOMBI are summarized by Table 1.

2.2 Conference

The Conference track contains 16 ontologies from the same domain (conference organization). They have been developed within the OntoFarm project[2]. This

[1] http://oaei.ontologymatching.org/2015/tutorial/tutorialv4.pdf
[2] http://owl.vse.cz:8080/ontofarm/

R.A.M.	Precision	F0.5 Measure	F1 Measure	F2 Measure	Recall
ra1-M1	0.64	0.64	0.64	0.65	0.65
ra1-M2	0.29	0.27	0.24	0.21	0.2
ra1-M3	0.61	0.6	0.59	0.59	0.58
ra2-M1	0.6	0.6	0.59	0.58	0.58
ra2-M2	0.29	0.27	0.24	0.21	0.2
ra2-M3	0.57	0.56	0.55	0.54	0.53
rar2-M1	0.59	0.59	0.6	0.61	0.61
rar2-M2	0.29	0.27	0.24	0.21	0.2
rar2-M3	0.56	0.56	0.55	0.55	0.55

Table 2: Results of ServOMBI for the Conference track.

Task	Precision	Recall	F-Measure
FMA-NCI	0.97	0.806	0.88
FMA-SNOMED	0.96	0.664	0.785

Table 3: Initial results of ServOMBI for the Large Bio track.

year the different tools are evaluated using i) crisp reference alignments where the confidence values for all matches are 1.0, ii) the uncertain version of the reference alignment where confidence values reflect the degree of agreement of a group of twenty people on the validity of the match [6] and iii) logical reasoning using violations of consistency and conservativity principles [15] [16]. Various reference alignments and evaluation modalities (R.A.M.) are used to assess the performance of the tooms. Thus, ra1 is the original reference alignment of the Conference track, ra2 is entailed reference alignment generated as a transitive closure computed on the original reference alignment (ra1) and rar2 is violation free version of reference alignment. Three different modalities are provided for these reference alignments, M1, M2 and M3 which contain respectively only classes, only properties and classes and properties.

The results obtained by ServOMBI according to these different modalities on the crisp reference alignments where the confidence value is 1.0 are summarized on table 2. The value of β is respectively set to 0.5, 1 (harmonic measure) and 2.

2.3 Largebio

The Large BioMed track consists of finding alignments between the Foundational Model of Anatomy (FMA), SNOMED CT, and the National Cancer Institute Thesaurus (NCI). The results obtained by ServOMBI for the small fragments of FMA-NCI task and FMA-SNOMED ontologies are summarize in Table 3

2.4 Interactive track

This track aims at offering a systematic and automated evaluation of matching systems with user interaction to compare the quality of interactive matching approaches in terms of F-measure and number of required interactions. For the 2015 edition, the Conference, Anatomy and Largebio tracks dataset are used

Error rate	Precision	Recall	F-Measure
0.0	1.00	0.617	0.763
0.1	1.00	0.587	0.740
0.2	1.00	0.553	0.712
0.3	1.00	0.519	0.683

Table 4: Results of ServOMBI for the Interactive track on the Anatomy dataset.

Error rate	Precision	Recall	F-Measure
0.0	1.00	0.650	0.788
0.1	1.00	0.637	0.778
0.2	1.00	0.622	0.767
0.3	1.00	0.627	0.770

Table 5: Results of ServOMBI for the Interactive track on the Conference dataset.

for the evaluation. Moreover, this year a domain experts with variable error rates, respectively 0.1, 0.2 and 0.3 are considered in addition to the perferct emmulated user (oracle) with error rate 0.0. ServOMBI participated for the first year to this track. The interaction implemented currently in the system is mainly to allow the user validating the provided candidate mappings. Tables 4, 5 and 6 give respectively the results obtained by the system on the Anatomy, Conference and Largebio dataset for the Interactive track. We note that for the Largebio interactive track, the ServOMBI was only able to match the FMA-NCI small fragments and FMA-SNOMED small fragments.

Overall ServOMBI improved its performance when compared to the results obtained with the normal Anatomy, Conference and Largebio track. However, the system make a greater number of requests compared to the other participating systems in the Interactive track.

2.5 Ontology Alignment for Query Answering

This track does not follow the usual OAEI tasks for evaluating the performance of participating systems [14]. Precision and Recall are calculated with respect to the ability of the generated alignments to answer a set of queries in a ontology-based data access scenario where several ontologies exist. This track uses the Conference dataset for the evaluation with two reference alignments, the publicly available Conference track alignment (RA1) and the repaired one (RAR1). Table 7 summarizes the results of ServOMBI which succeed with 6 out of 18 queries.

Error rate	Precision	Recall	F-Measure
0.0	1.00	0.737	0.847
0.1	1.00	0.716	0.832
0.2	1.00	0.688	0.813
0.3	1.00	0.660	0.792

Table 6: Results of ServOMBI for the Interactive track on the Largebio dataset.

Task	Precision	Recall	F-Measure
OAQA RA1	0.222	0.222	0.222
OAQA RAR1	0.222	0.222	0.222

Table 7: Results of ServOMBI for the OAQA track.

3 General Comments

We have participated in the 2012 and 2013 edition with the ServOMap system which achieved overall good results. The performance of this system is very good in particular for the tasks involving large ontologies. The new features implemented within ServOMBI did not lead to overall improved performances according to the results of the ServOMap system as expected. The contextual similarity computation, which is performed iteratively, is very time consuming and did not improved the overall recall of the system. In addition, while there is a gain in terms of computation times with concepts lookups, the BOI does not impact the overall performance of the system in terms of times taken to perform the matching tasks.

4 Conclusion

We have described in this paper the main functionalities of the ServOMBI ontology matching system and the overall results obtained during the 2015 OAEI edition. ServOMBI introduces a binary indexing strategy to complement the usual terminological indexing strategy used by the ServOMap system. The system achieved performance lower than expected according to the introduced features for the contextual similarity coputation. However it succed improving the F-Measure whith the interaction strategy. ServOMap continues to be developed in parralel and now include graph-based visualization.

As of future work, we envision to investigate an improved integration of the binary indexing and the contextual similarity computing. In addition, we plan to use combine multiple learning algorithms to improve the candidate selection during the contextual similarity computing.

5 Acknowledgments

This work has been partly supported by the project DRUG-SAFE funded ANSM (Agence Nationale de la Sécurité du Médicament). We also thank the organizers of OAEI with providing test dataset and the evaluation infrastructure.

References

[1] Jerôme Euzenat and Pavel Shvaiko, "Ontology Matching", Springer-Verlag, Heidelberg, 2013

[2] Gayo, Diallo. An effective method of large scale ontology matching. Journal of Biomedical Semantics, vol:5(44), 2014. DOI:10.1186/2041-1480-5-44

[3] Gayo Diallo. Efficient building of local repository of distributed ontologies. IEEE Proc. of the 7th International Conference on Signal Image Technology & Internet Based Systems (SITIS'11). K Yetongnon, R Chbeir and A Dipanda eds. Nov 28- Dec 1st 2011, Dijon, France

[4] Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Modern Information Retrieval—The Concepts and Technology behind Search. 2nd Edition, Pearson, 2011

[5] Samir Amir, Hassan Aït-Kaci. CEDAR: Efficient Reasoning for the Semantic Web. Proceedings of the 10th IEEE International Conference on Signal Image Technology & Internet-Based Systems (SITIS 2014), Marrakech, Morocco, 2014

[6] Michelle Cheatham, Pascal Hitzler. Conference v2.0: An Uncertain Version of the OAEI Conference Benchmark. International Semantic Web Conference (2) 2014: 33-48

[7] George A. Miller. Wordnet: A lexical database for english. Communications Of The ACM, 38:39–41, 1995.

[8] Iwama, Kazuo; Miyazaki, Shuichi. A Survey of the Stable Marriage Problem and Its Variants. pp. 131–136. doi:10.1109/ICKS.2008.7

[9] Yevgeny Kazakov, Markus Krötzsch, František Simančík. Unchain My EL Reasoner. In Riccardo Rosati, Sebastian Rudolph, Michael Zakharyaschev, eds.: Proceedings of the 24th International Workshop on Description Logics (DL-11). CEUR Workshop Proceedings 2011

[10] Niels Landwehr, Mark Hall, and Eibe Frank. Logistic Model Trees . In Machine Learning 59 (1-2) 161-205, 2005

[11] Birte Glimm, Ian Horrocks, Boris Motik , Giorgos Stoilos, Zhe Wang. HermiT: An OWL 2 Reasoner. Journal of Automated Reasoning. Volume 53, Issue 3, pp 245-269, 2014

[12] Heiko Paulheim, Sven Hertling, Dominique Ritze. "Towards Evaluating Interactive Ontology Matching Tools". ESWC 2013

[13] Gerard Salton. (1979). Mathematics and information retrieval. Journal of Documentation,35 (),–29

[14] Alessandro Solimando, Ernesto Jimenez-Ruiz, and Christoph Pinkel. Evaluating Ontology Alignment Systems in Query Answering Tasks. Poster paper at International Semantic Web Conference (ISWC). 2014

[15] Alessandro Solimando, Ernesto Jiménez-Ruiz, Giovanna Guerrini. Detecting and Correcting Conservativity Principle Violations in Ontology-to-Ontology Mappings. International Semantic Web Conference (2) 2014: 1-16.

[16] Alessandro Solimando, Ernesto Jiménez-Ruiz, Giovanna Guerrini. A Multistrategy Approach for Detecting and Correcting Conservativity Principle Violations in Ontology Alignments. OWL: Experiences and Directions Workshop 2014 (OWLED 2014). 13-24