

LYAM++ Results for OAEI 2015

Abdel Nasser Tigrine, Zohra Bellahsene, Konstantin Todorov

► **To cite this version:**

Abdel Nasser Tigrine, Zohra Bellahsene, Konstantin Todorov. LYAM++ Results for OAEI 2015. OM: Ontology Matching, Oct 2015, Bethlehem, PA, United States. 10th International Workshop on Ontology Matching collocated with the 14th International Semantic Web Conference, 2015, <<http://www.om2015.ontologymatching.org>>. <lirmm-01408026>

HAL Id: lirmm-01408026

<https://hal-lirmm.ccsd.cnrs.fr/lirmm-01408026>

Submitted on 2 Dec 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LYAM++ Results for OAEI 2015

Abdel Nasser Tigrine, Zohra Bellahsene, Konstantin Todorov

{lastname}@lirmm.fr
LIRMM / University of Montpellier, France

Abstract. The paper presents a novel technique for aligning cross-lingual ontologies that does not rely on machine translation, but uses the large multilingual semantic network BabelNet as a source of background knowledge. In addition, our approach applies a novel orchestration of the components of the matching workflow. We present our results on the evaluation challenge Multifarm.

1 Presentation of the System

In spite of the considerable advance that has been made in the field of ontology matching recently, many questions remain open [1]. The current work addresses the challenge of using background knowledge with a focus on aligning cross-lingual ontologies, i.e., ontologies defined in different natural languages [2].

Indeed, considering multilingual and cross-lingual information is becoming more and more important, in view particularly of the growing number of web content-creating non-English users and the clear demand of cross-language interoperability. In the context of the web of data, it is important to propose procedures for linking vocabularies across natural languages, in order to foster the creation of a veritable global information network.

The use of different natural languages in the concepts and relations labeling process is becoming an important source of ontology heterogeneity. The methods that have been proposed to deal with it most commonly rely on automatic translation of labels to a single target language [3] or apply machine learning techniques [2]. However, machine translation tolerates low precision levels and machine learning methods require large training corpus that is rarely available in an ontology matching scenario. An inherent problem of translation is that there is often a lack of exact one-to-one correspondence between the terms in different natural languages.

1.1 State, Purpose, General Statement

We present LYAM++ (Yet Another Matcher - Light)[4], a fully automatic cross-lingual ontology matching system that does not rely on machine translation. Instead, we make use of the openly available general-purpose multilingual semantic network BabelNet¹ in order to recreate the missing semantic context

¹ <http://babelnet.org/>

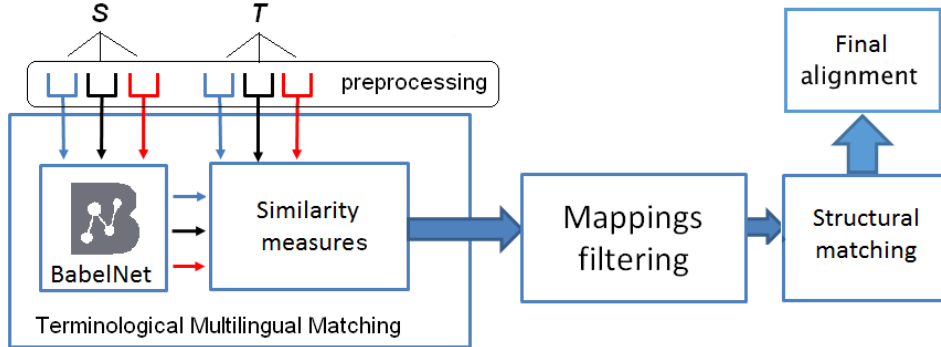


Fig. 1: The processing pipeline of LYAM++.

in the matching process. Another original feature of our approach is the choice of orchestration of the matching workflow. Our experiments on the MultiFarm² benchmark data show that (1) our method outperforms the best approaches in the current state-of-the-art and (2) the novel workflow orchestration provides better results compared to the classical one. We refer the reader to the results reported in [4].

1.2 Specific Techniques Used

The workflow of LYAM++ is given in Fig. 1. We take as an input a source ontology S , given in a natural language l_S and a target ontology T , given in a language l_T . The overall process consists of four main components: a terminological multilingual matcher, a mapping selection module and, finally, a structural matcher. One of the original contributions of this work is the choice of orchestration of these components. Indeed, the places of the mapping selection module and the structural matcher are reversed in the existing OM tools [5]. However, we wanted to ensure that we feed only good quality mappings to the structural matcher, therefore we decided to filter the discovered correspondences right after producing the initial alignment. This decision is supported experimentally in the following section.

The *multilingual terminological matching* module, the second contribution described in this paper, acts on the one hand as a preprocessing component and, on the other hand – as a light-weight terminological matcher between cross-lingual labels. We start by splitting the elements of each ontology in three groups: labels of classes, labels of object properties and labels of data object properties (in colors blue, black and red in the figure), since these groups of elements are to be aligned separately. A standard preprocessing procedure is applied on these sets of labels, comprising character normalization, stop-words filtering, tokenization and lemmatization. The tokens of the elements of T are then aligned

² <http://web.informatik.uni-mannheim.de/multifarm/>

to BabelNet. At first, every token of a given label s in S is enriched by related terms and synonyms from BabelNet and all of these terms are represented in the language l_T , which makes these terms comparable to the tokens of the labels in T . A simple similarity evaluation by the help of the Jaccard coefficient selects the term in each set of related terms corresponding to a given token from s that has the highest score with respect to every token in each label of T . This helps to reconstitute the label s in the language l_T . Finally, the labels in each group of S and T , seen as sets of tokens, are compared by using the Soft TFIDF similarity measure [6], which produces an intermediate terminological alignment.

The three remaining components are standard OM modules [5], although ordered in a new manner. The *Mapping selection* is a module that transforms the initial 1 to many mapping to a 1:1 mapping based on the principle of iteratively retaining the pairs of concepts with maximal value of similarity. Finally, the *structural matcher* component filters the trustworthy pairs of aligned concepts by looking at the similarity values produced for their parents and their children in the ontology hierarchies.

1.3 Links to the System and to the Set of Provided Alignments

The system is not yet available online because it depends heavily on the use of BabelNet, which is under a non-free licence. We are working on implementing a sharable version of LYAM++ making use of different open access background knowledge sources, such as YAGO.

The alignments produced by LYAM++ for this year’s Multifarm track can be found under the following link: <http://www.lirmm.fr/benellefi/Lyam++.rar>

2 Results

We have evaluated our approach on data coming from the ontology alignment evaluation initiative (OAEI)³ and particularly Multifarm—a benchmark designed for evaluating cross-lingual ontology matching systems. Multifarm data consist of a set of 7 ontologies originally coming from the *Conference* benchmark of OAEI, translated into 8 languages. Two evaluation tasks are defined: *task 1* consists in matching two different ontologies given in different languages, while *task 2* aims to align different language versions of one single ontology.

We have performed experiments on both tasks by using the pairs of languages given in the summary of our results in Table 1.

In another experiment, we have evaluated the results obtained by using our novel orchestration of matching components, as compared to the standard orchestration. The figures in Table 2 show that the workflow proposed in this paper acts in favor of achieving better results as compared to the standard method.

Table 3 presents the results obtained by LYAM++ on this year’s Multifarm evaluation campaign. What we see is the average F-measure value for all

³ <http://oaei.ontologymatching.org/>

Table 1: Results of LYAM++ on the Multifarm datasets.

Lang. pair	FR-RU	FR-PT	FR-NL	ES-FR	ES-RU	ES-PT	ES-NL	EN-PT	EN-RU	EN-FR
LYAM++	0.54	0.58	0.62	0.60	0.60	0.60	0.63	0.67	0.53	0.59

Average F-measures over all threshold values per language pair for task 1.

Lang. pair	FR-RU	FR-PT	FR-NL	ES-FR	ES-RU	ES-PT	ES-NL	EN-PT	EN-RU	EN-FR
LYAM++	0.58	0.72	0.67	0.77	0.64	0.70	0.68	0.74	0.59	0.85

Average F-measures over all threshold values per language pair for task 2.

Table 2: Comparing the standard and the novel orchestrations

Language pair	FR-RU	FR-PT	FR-NL	ES-FR	ES-RU	ES-PT	ES-NL	EN-PT	EN-RU	EN-FR
Novel	0.58	0.72	0.67	0.77	0.64	0.70	0.68	0.74	0.59	0.85
Standard	0.50	0.58	0.60	0.39	0.54	0.57	0.58	0.50	0.32	0.39

Average F-measures over all threshold values per language pair for task 2.

language-pairs without any threshold on the confidence measure. The value in the parenthesis corresponds to the average F-measure value for the generated alignments only (the pairs of languages that the system handles). Note that at this stage, LYAM++ is not able to generate alignments for all language-pairs because of query constraints related to the use of BabelNet, which is under non-free license. The table compares LYAM++ to the top 3 systems this year (LYAM++ excluded). We see that with respect to the languages that our system can handle, it scores second, little behind AML, on task 1 and first on task 2.

Table 3: Results of LYAM++ for Multifarm 2015.

	Task1	Task2
LYAM++	0.14(0.49)	0.19(0.66)
AML	0.51(0.51)	0.64(0.64)
LogMap	0.41(0.41)	0.45(0.45)
CLONA	0.39(0.39)	0.58(0.58)

3 Possible Improvements of the System

Currently, we are working on enhancing the system in order to make it applicable to the general ontology matching problem and not only to the cross-lingual one. We have generated first results on the Conference benchmark without any modification in the settings and our results are promising. For the majority of the datasets (ontology pairs) our system achieves a F-score almost as good as the F-score of AML, the best performing system on that track.

We consider that a key feature for the improvement of our system is the appropriate choice of background knowledge. In order to improve the results achieved on the Conference track, we plan to use monolingual general purpose background knowledge (for example, the english subgraphs of YAGO or DBPedia) instead of BabelNet.

We intend to use domain specific background knowledge in order to solve alignment problems in specific areas of knowledge. More precisely, we plan to participate on the Anatomy track by testing different kinds of domain specific background knowledge, such as UMLS or other.

4 Conclusion

We presented an efficient approach for aligning cross-lingual ontologies by using the multilingual lexical database BabelNet. Subjects of ongoing and future work are (1) testing and evaluating different sources of external knowledge, (2) applying the approach to a larger set of languages, (3) adaptation of the approach to the monolingual case and studying the use of background knowledge in a monolingual ontology matching scenario.

References

1. P. Shvaiko and J. Euzenat, “Ontology matching: state of the art and future challenges,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, no. 1, pp. 158–176, 2013.
2. D. Spohr, L. Hollink, and P. Cimiano, “A machine learning approach to multilingual and cross-lingual ontology matching,” in *The Semantic Web–ISWC 2011*, pp. 665–680, Springer, 2011.
3. D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, and F. M. Couto, “The agreementmakerlight ontology matching system,” in *On the Move to Meaningful Internet Systems: OTM 2013 Conferences*, pp. 527–541, Springer, 2013.
4. A. N. Tigrine, Z. Bellahsene, and K. Todorov, “Light-weight cross-lingual ontology matching with LYAM++,” in *On the Move to Meaningful Internet Systems: OTM 2015 Conferences - Confederated International Conferences: CoopIS, ODBASE, and C&TC 2015, Rhodes, Greece, October 26-30, 2015, Proceedings*, pp. 527–544, 2015.
5. D. Ngo, Z. Bellahsene, and K. Todorov, “Opening the black box of ontology matching,” in *The Semantic Web: Semantics and Big Data*, pp. 16–30, Springer, 2013.
6. W. W. Cohen, P. D. Ravikumar, and S. E. Fienberg, “A comparison of string distance metrics for name-matching tasks,” in *IJWeb*, pp. 73–78, 2003.