YAM-BIO: Results for OAEI 2017

Amina Annane[1,2], Zohra Bellahsene[1], Faical Azouaou[2], and Clement Jonquet[1,3]

[1] University of Montpellier LIRMM, France
lastname@lirmm.fr
[2] National Higher School of Computer Science (ESI), Algeria
[3] Center for Biomedical Informatics Research, Stanford University, USA

Abstract. YAM-BIO system is an extension of YAM++ system, which is dedicated to match biomedical ontologies. This extension includes a new component that uses existing mappings as background knowledge. In this paper, we present YAM-BIO and its results that are obtained in Anatomy and largebio tracks of OAEI 2017 campaign.

1 Presentation of the system

1.1 State, purpose, general statement

YAM-BIO can be seen as an extension of YAM++ [4] with the use of existing mappings as background knowledge to enhance the matching of biomedical ontologies. The version of YAM++, that we reuse in YAM-BIO, got excellent results in OAEI 2013 campaign [8], since YAM++ did not participate more. Four years on from the last participation, we aim to establish a comparison between the performance of YAM++ and state-of-the-art systems in matching biomedical ontologies. In the last OAEI campaigns, state-of-the-art systems such as AML [6] and LogMapBio [7] used specialized background knowledge to improve their results. To make a fair comparison, we added a layer that uses existing mappings as background knowledge. This year, YAM-BIO participates in two tracks: Anatomy and Large biomedical ontologies.

1.2 Specific techniques used

As we can see in Fig. 1, YAM-BIO workflow contains three main steps. The first one consists in discovering the direct matching between the source and the target ontologies by using YAM++. The second step tries to find mappings for source concepts that have not been matched by composing existing mappings. The third step consists in processing the union of the alignments produced by the previous steps.

Direct matching with YAM++. The matching process starts by loading ontologies using the ontology loader. Then, annotations and structure of source

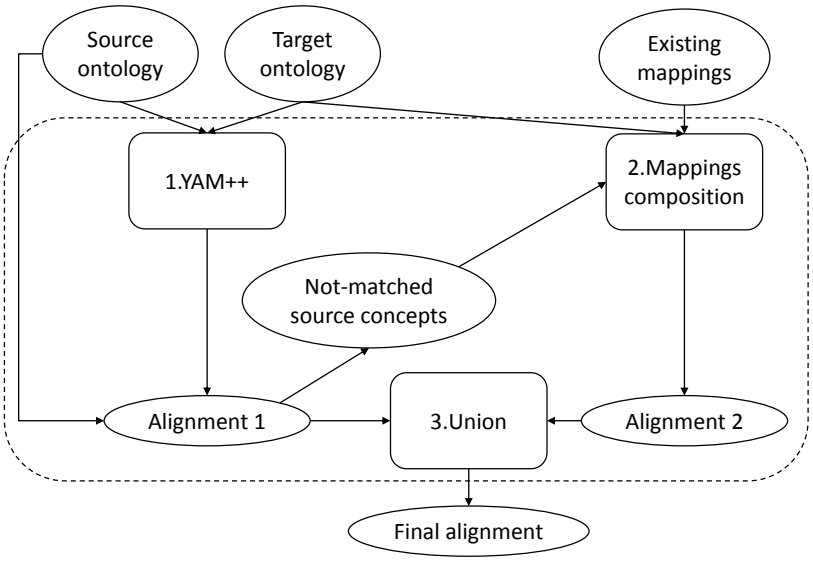

Fig. 1. YAM-BIO: general workflow

and target ontologies are indexed as well as the context of each entity. The candidate pre-filtering module eliminates all candidate mappings that have a low annotation similarity. Other advanced lexical and structural similarity measures are applied on the remaining candidate mappings during the similarity computation step. Then, the similarity propagation step updates similarity scores of candidate mappings using structures of source and target ontologies. A threshold is dynamically computed to select the most relevant mapping candidates during the candidate post-filtering step. For more details, we refer readers to the YAM++ overview paper [4].

Indirect matching. The aim of this step is to find mappings for the concepts that have not been matched by the direct matching. First, the existing mappings that form our background knowledge are loaded in a list of lists called A as follows:

1. Codes of all concepts present in the background knowledge are added to A.
2. Each element x of A points a list that contains codes of all concepts matched to x in the background knowledge.

Then, for each source concept y that is not matched yet by the first component, we check if y code exists in A. If this is the case, we get the list pointed by y and for each element of this list we verify if it points to a list that contains a concept code that belongs to the target ontology. If so, we derive a new mapping that we add to the alignment produced previously by the direct matching.

1.3 Adaptations made for the evaluation

The existing mappings used as background knowledge have been extracted from UBERON and DOID ontologies. Indeed, these ontologies contains several cross references to other ontologies that may be considered as manual mappings. In addition, concept codes of large biomedical ontologies are not the original ones, but they have been replaced by their preferred labels in the large biomedical ontologies dataset. For that, we have implemented a method that used the REST API of NCBO BioPortal [5] to replace the codes of referenced concepts in UBERON and DOID by their preferred labels.

1.4 Link to the system

The YAM++ system has an online version [10] at the link http://yamplusplus.lirmm.fr/

1.5 Link to the set of provided alignments

The set of produced alignments as well as the background knowledge file are available at the following link https://goo.gl/zNznNz

2 Results

2.1 Anatomy

The Anatomy track consists of finding an alignment between the Adult Mouse Anatomy (2744 classes) and a part of the NCI Thesaurus (3304 classes) describing the human anatomy. Table 1 shows the evaluation result and runtime of YAM-BIO on this track. YAM-BIO has the second position in Anatomy track

Test set	Precision	Recall	F-Score	Time (s)
Anatomy	0.948	0.922	0.935	70

Table 1. Anatomy results

among the 12 systems that have participated with almost the same precision and lower recall comparing to the top ranked system.

2.2 Large Biomedical Ontologies

This track consists of finding alignments between the Foundational Model of Anatomy (FMA), SNOMED CT, and the National Cancer Institute Thesaurus (NCI). There are six tasks with different size of input ontologies, i.e., small fragment, large fragment and the whole ontologies. Table 2 shows the evaluation results and run times of YAM-BIO on those tasks.

Test set	Precision	Recall	F-Score	Time (s)
Task 1: Small FMA-NCI	0.968	0.896	0.931	56
Task 2: Whole FMA-NCI	0.816	0.888	0.850	279
Task 3: Small FMA-SNOMED	0.966	0.733	0.834	60
Task 4: Whole FMA-SNOMED	0.887	0.728	0.800	468
Task 5: Small SNOMED-NCI	0.899	0.677	0.772	2202
Task 6: Whole SNOMED-NCI	0.827	0.698	0.757	490

Table 2. LargeBio results

Without considering the XMAP system[4], YAM-BIO is the top ranked system in Task 1 and Task 4 and it has almost the same result as the best system in Task 3 with an F-measure of 0.834 vs 0.835 that is obtained by the best system. In Task 2 and Task 6, YAM-BIO has the second position with a better recall than the best system and a lower precision. In Task 5, it shares the third position with LogMapBio. In terms of running time, YAM-BIO completed the different tasks in acceptable time, except for Task 5

3 General comments

3.1 Comments on the results

YAM-BIO got the second position in Anatomy track and its results in the Large biomedical ontologies track are close the one of the competing systems. Indeed, YAM-BIO is among the two top ranked systems in all tasks except task 5 of the LargeBio track where it shares the third position with LogMapBio. The use of the existing mappings as background knowledge has improved the results in terms of recall and F-measure. The composition of the mappings extracted from UBERON ontology allowed to discover non trivial mappings, specifically in Anatomy and in Task 1 and Task 2 of largebio track. In the same way the mappings extracted from DOID allowed to increase the recall of Task 5 and Task 6. While the good results of Task 3 and Task 4 reflect the effectiveness of information content matching techniques implemented in YAM++. However, the incoherence analysis shows that YAM-BIO presents a certain unsatisfiability of the discovered mappings. This may be explained by the fact that the mappings derived using background knowledge have been added to the final alignment without any semantic verification.

3.2 Discussions on the way to improve the proposed system

Currently, the mappings derived using background knowledge are not concerned by the post-filtering and semantic verification steps. The produced alignment is obtained by performing the union of the alignments produced by the direct and

[4] XMAP uses UMLS-Metathesaurus as background knowledge, the source from which largebio reference alignment is extracted.

indirect components. In the future we aim to integrate the use of background knowledge in the internal architecture of YAM++ which will improve coherence of the final results, more specifically we aim to implement our approach proposed in [1]. In addition, we are aware of the importance of the dynamic selection of ontologies to use them as background knowledge [9, 2]. Indeed, from the selected ontologies we may extract manual and automatic mappings to reuse them as background knowledge. For that, we aim to extend YAM-BIO with a novel layer that will select dynamically a set of ontologies from a generic repository such as Watson [3].

3.3 Comments on the evaluation

We think that it would be interesting to publish the results of the participants with and without the use of specialized background knowledge when it is possible (i.e, when the system can work without its BK). On the one hand, this will allow to better evaluate the use of BK in terms of matching quality and running time. On the other hand, it will enable a fair comparison with systems that do not use background knowledge.

Some components are common in the architecture of all ontology matching systems. For example the element level component is implemented in all systems even if they do not use the same similarity measures. Other components such as the background knowledge selection or the semantic verification are not available in all systems. We think that the comparison of the running time of whole systems is not fair, and it is better to evaluate the time taken by each component separately if possible. For example, YAM-BIO used a predefined background knowledge while LogMapBio made a dynamic selection from a distant repository. In this case, it is normal that LogMapBio took more time than YAM-BIO and any comparison between the two systems in terms of running time is unfair. For that, it would be interesting to know how long the selection has taken. Separating the running time of each component will also help the community to identify the components that are not efficient in order to improve them, and those that are efficient to reuse them.

4 Conclusion

In this year, YAM-BIO participated in two tracks Anatomy and LargeBio. The obtained results are very close to the top ranked state-of-the-art systems thanks to the different techniques of content matching implemented in YAM++ and the use of background knowledge. However, due to the high heterogeneity of ontologies, we believe that an advanced module that selects and uses background knowledge should be implemented in the internal architecture of YAM++ to improve its results. In the future, we will try to implement this module and participate in different tracks of OAEI.

5 Acknowledgment

The authors acknowledge the Eiffel Excellence Scholarship program.This work was done during a LIRMM-ESI collaboration within the Semantic Indexing of French biomedical Resources and PractiKPharma project that received funding from the French National Research Agency (grant ANR-12-JS02-01001 and ANR-15-CE23-0028) as well as by the European H2020 Marie Sklodowska-Curie action (agreement No 701771), the University of Montpellier and the CNRS.

References

1. Annane Amina, Bellahsene Zohra, Azouaou Faical, and Jonquet Clement. Selection and combination of heterogeneous mappings to enhance biomedical ontology matching. In *20th International Conference on Knowledge Engineering and Knowledge Management, EKAW, Bologna, Italy*, pages 19–33, 2016.
2. Faria Daniel, Pesquita Catia, Santos Emanuel, Cruz Isabel F, and Couto Francisco M. Automatic background knowledge selection for matching biomedical ontologies. *PloS One*, 9(11):e111226, 2014.
3. d'Aquin Mathieu, Gridinoc Laurian, Angeletou Sofia, Sabou Marta, and Motta Enrico. Watson: A Gateway for Next Generation Semantic Web Applications. In *6th International Semantic Web Conference, ISWC, Poster and Demonstration, Busan, Korea*, pages 11–15, 2007.
4. Ngo DuyHoa and Bellahsene Zohra. Overview of YAM++:(not) yet another matcher for ontology alignment task. *Journal of Web Semantics*, 41:30 – 49, 2016.
5. Noy Natalya F, Shah Nigam H, Whetzel Patricia L, Dai Benjamin, Dorf Michael, Griffith Nicholas, Jonquet Clement, Rubin Daniel L, Storey Margaret-Anne, and Chute Christopher G. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37:170–173, 2009.
6. Daniel Faria, Catia Pesquita, Booma S Balasubramani, Catarina Martins, Joao Cardoso, Hugo Curado, Francisco M Couto, and Isabel F Cruz. Oaei 2016 results of AML. In *11th International Workshop on Ontology Matching, Kobe, Japan.*, pages 138–145, 2016.
7. E Jiménez-Ruiz, B Cuenca Grau, and V Cross. Logmap family participation in the oaei 2016. In *11th International Workshop on Ontology Matching, Kobe, Japan.*, pages 185–189, 2016.
8. DuyHoa Ngo and Zohra Bellahsene. YAM++ results for OAEI 2013. In *8th International Workshop on Ontology Matching, Sydney, Australia.*, pages 211–218, 2013.
9. Chen Xi, Xia Weiguo, Jiménez-Ruiz Ernesto, and Cross Valerie. Extending an ontology alignment system with BioPortal: a preliminary analysis. In *13th International Semantic Web Conference, ISWC, Posters and Demonstrations, Riva del Garda, Italy*, pages 313–316, 2014.
10. Bellahsene Zohra, Emonet Vincent, Ngo DuyHoa, and Todorov Konstantin. Yam++ online: a multi-task platform for ontology and thesaurus matching. In *14th Extended Semantic Web Conference, ESWC, Posters and Demonstrations, Portoroz, Slovenia*, 2017.