CEUR-WS.org/Vol-2788/om2020_STpaper1.pdf

Learning reference alignments for ontology matching within and across domains [*]

Beatriz Lima[1], Ruben Branco[2,3], João Castanheira, Gustavo Fonseca[1], and Catia Pesquita[1,3]

[1] LASIGE, Faculdade de Ciências da Universidade de Lisboa, Portugal
[2] NLX—Natural Language and Speech Group, Faculdade de Ciências da Universidade de Lisboa, Portugal
[3] Dep. de Informática, Faculdade de Ciências da Universidade de Lisboa, Portugal
clpesquita@fc.ul.pt

Abstract. Reference alignments are the standard approach for ontology alignment evaluation. However, building a reference alignment is time-consuming and usually depends on expert availability. Several strategies have been proposed to mitigate this issue, ranging from exploring external resources, building simulated alignment tasks, or even crowdsourcing. A simple approach is to take a consensus alignment built from the outputs of several ontology matching systems results.

We present a preliminary investigation that focuses on the generalization of machine learning models trained on the output alignments of multiple systems for a task where a reference alignment is available to other alignment tasks.

Results show that while the consensus alignment works well for alignment tasks where several systems achieve a high performance and produce similar alignments, trained reference models are able to improve on the consensus both within and across domains.

Keywords: Ontology matching · Machine Learning · Reference alignment · OAEI

1 Introduction

The evaluation of ontology alignments typically relies on reference alignments which are automatically compared to the outputs of the alignment systems. Reference alignments are commonly either manually-curated by domain experts or automatically generated. The first kind can be created manually from scratch or manually validated given a set of automatically generated candidates[10]. Although very reliable, they are difficult to obtain as they are very time-consuming and require domain expertise. To decrease the effort and associated cost, both automated strategies and crowdsourcing have been used. Automated strategies, usually work with simulated data [5] or by exploring external resources [9].

[*] Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Crowdsourcing has been successfully employed, however producing references for complex domains is more difficult to achieve due to the lack of expertise of crowdsourced workers[2]. When the above options are not available, an easy solution to evaluate competing systems is based on a consensus alignment. This strategy is employed by the Disease and Phenotype track at the Ontology Alignment Evaluation Initiative [7] with a consensus alignment built on three votes (i.e., if a mapping is found by 3 different systems it is considered correct). The consensus is considered to be a partial reference alignment and mappings that are generated by a single system are then manually evaluated.

Motivated by the difficulties in generating a reference alignment and inspired by the consensus alignment strategy, we hypothesise that a machine learning model trained on the output alignments of multiple systems for a task where a reference alignment is available, can be used to evaluate other alignment tasks.

2 Methodology

The alignments produced by the ontology matching (OM) tools that participated in the Anatomy, Large BioMed and Conference tracks of OAEI 2019[4] were used as data sources. In our proposed models, each instance corresponds to a mapping in a given alignment task. The features translate in whether the given mapping was present or absent in the output of each of the participating OM tools, taking as values 1 or 0. The reference alignment was used to produce the target class, and support supervised learning. Thus, the model is learning to classify whether a mapping between two ontologies is correct, based on the pattern of outputs of the OM tools while using the reliable reference alignment as ground-truth.

The Anatomy track consists of matching Adult Mouse Anatomy[8] and the portion covering human anatomy of the National Cancer Institute Thesaurus (NCI)[6], and it is supported by a manually-curated reference alignment. Several OM systems achieve a high performance in this track [4]. The Large BioMed track comprises three ontologies, the Foundational Model of Anatomy (FMA)[11], SNOMED CT[3], and NCI. These ontologies are matched pairwise, generating three possible alignments: FMA-NCI, FMA-SNOMED and NCI-SNOMED, which will be further addressed as LB1, LB2 and LB3, respectively, for abbreviation. LB1 and LB2 cover the anatomical domain, whereas LB3 does not. The reference alignment was extracted from an external resource [9]. The Conference track[13] provides 16 ontologies from the conference organisation domain. Since only 7 ontologies are contained in the existing reference alignment, we end up with 21 result alignments, which corresponds to the complete alignment space between these ontologies. We randomly generated 3 different datasets (CF1, CF2, CF3), each of which containing 18 alignments for training and 3 alignments for testing. The alignments used for testing were cmt-ekaw, cmt-conference and iasted-sigkdd in CF1; conference-confof, edas-sigkdd and iasted-sigkdd in CF2; confof-edas, cmt-ekaw and ekaw-sigkdd in CF3.

[4] http://oaei.ontologymatching.org/2019/

A reference alignment only contains true positive mappings. Assuming its completeness, every potential mapping that is not a part of the reference is a false mapping. A traditional option to generate negative examples would be a random sampling of entity pairs from each ontology that are not present in the reference alignment. However, this would result in mostly instances with all zero features, and thus uninformative, since most systems produce alignments of cardinality near one. Instead, we take as negative examples all mappings that at least one of the OM tools finds but which are not a part of the reference alignment. To tackle the imbalance caused by this approach, we investigated two different sampling strategies: SMOTE oversampling[1] and undersampling with TomekLinks[12].

Three types of experiments were performed for each domain to verify different properties. In **Experiment 1**, which worked as a baseline, we investigated how well a model can be learned within a given alignment task. We performed 10-fold cross-validation, with a grid search for hyperparameter tuning over a set of 8 machine learning approaches[5]. In **Experiment 2**, we investigated if a model trained in one/more tasks would generalize well to other tasks within the same domain. To support this, features were extracted from the OM tools which participated in both training and test tasks. **Experiment 3** aims to verify how well the method generalises for ontologies in completely different domains. We train on LargeBio data and test on Conference, and vice-versa, again using the intersection of OM tools that participated in both tracks. For all experiments, we also computed the majority vote and the consensus with vote=3 results.

3 Results and Discussion

Table 1 presents the results obtained for all three experiments, using the best sampling strategy (oversampling) and machine learning approaches [6]. In the Biomedical domain, all cross-validation experiments achieved good performance (0.8 to 0.915 average F1-score), however, in the Anatomy task, the *Three votes* approach achieved the best result. In the second experiment, the model learned in Anatomy achieved at best an F1-score of 0.697 in LargeBio, whereas the model trained on LargeBio reached 0.938 in Anatomy. Nevertheless, the *Three votes* consensus approach achieved a higher score in these two cases. However, within the LB track, the ML models outperformed the consensus approach in LB1 and LB2 trained models. These results indicate that system strategies likely differ between the Anatomy and LargeBio tracks. The greater complexity and coverage of LB (which includes both anatomical and non-anatomical tasks) can help explain these results. In the Conference domain, the first experiment results were overall high, with ML approaches improving over the consensus. The second experiment revealed that the ML approaches were able to outperform

[5] Random Forest, K-Nearest Neighbors, Decision Tree, Multi-Layer Perceptron, Naive Bayes, Gradient Boosting, Logistic Regression and Adaboost

[6] The full table of results along with hyperparameter information can be found here: `https://github.com/liseda-lab/ML4ReferenceAlignment`

the consensus in only one test case. As for the cross-domain experiments, we can observe that, even though the LB dataset is much bigger than CF, models trained in CF were able to generalise well to LB and vice-versa, and in both cases surpass the consensus. One relevant aspect that may help explain these results is the agreement degree between OM systems. In the Anatomy task, the average agreement [7] between systems is 0.75, whereas in LB1, LB2, and LB3 it is 0.35, 0.26 and 0.40, respectively. In Conference the agreement ranges between 0.51 and 0.86 with most tasks falling below 0.65. This indicates that when systems have a high agreement, the consensus provides a good evaluation, but when systems differ in their outputs, the ML approaches work best.

Exp.	**Train**	**Test**	**Gradient Boosting**	**AdaBoost**	**Logistic Regression**	**Decision Tree**	Majority Vote	Three votes
				Biomedical				
1	Anatomy		0.915	0.897	0.902	0.909	0.907	**0.945**
	LB1		0.933	**0.935**	0.934	0.923	0.856	0.815
	LB2		**0.885**	0.806	0.822	0.881	0.384	0.689
	LB3		**0.905**	0.901	0.898	0.889	0.771	0.794
2	Anatomy	LB	0.376	0.697	0.629	0.380	0.712	**0.772**
	LB	Anatomy	0.937	0.860	0.147	0.938	0.907	**0.945**
	LB1	LB2+3	**0.794**	0.773	0.775	0.771	0.765	0.688
	LB2	LB1+3	**0.848**	0.807	0.834	0.836	0.786	0.798
	LB3	LB1+2	0.709	0.707	0.711	0.713	0.587	**0.734**
				Conference				
1	CF		**0.803**	0.767	0.770	0.766	0.616	0.668
2	CF1		0.682	**0.789**	0.771	0.651	0.585	0.510
	CF2		0.696	0.689	0.698	0.587	0.677	**0.720**
	CF3		0.571	0.634	0.634	0.609	**0.643**	0.635
				Cross-domain				
3	LB	CF	0.677	**0.678**	0.670	0.674	0.603	0.603
	CF	LB	0.783	0.787	**0.790**	0.788	0.720	0.720

Table 1. F1-scores using oversampling strategy and best classifiers. The values in bold are the best scoring classifiers for each experiment (row).

Our preliminary results highlight an opportunity to address the challenge of incomplete reference alignments by training models with a partial reference. Furthermore, they also showcase that in tasks where systems output dissimilar

[7] computed as the average pairwise jaccard similarity between OM systems outputs

alignments, a model trained in other alignment tasks, even from a different domain, can provide a more complete evaluation than a consensus alignment.

Acknowledgements CP and BL are funded by the FCT through LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020, and by projects SMILAX (ref. PTDC/EEI-ESS/4633/2014). CP is also funded by GADgET (ref. DSAIPA/DS/0022/2018). RB is funded by PORTULAN CLARIN Research Infrastructure through Lisboa 2020, Alentejo 2020 and FCT (PINFRA/22117/2016).

References

1. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
2. Cheatham, M., Hitzler, P.: Conference v2. 0: An uncertain version of the oaei conference benchmark. In: International Semantic Web Conference. pp. 33–48. Springer (2014)
3. Donnelly, K.: Snomed-ct: The advanced terminology and coding system for ehealth. Studies in health technology and informatics **121**, 279 (2006)
4. Dragisic, Z., Ivanova, V., Li, H., Lambrix, P.: Experiences from the anatomy track in the ontology alignment evaluation initiative. Journal of biomedical semantics **8**(1), 56 (2017)
5. Ferrara, A., Montanelli, S., Noessner, J., Stuckenschmidt, H.: Benchmarking matching applications on the semantic web. In: Extended Semantic Web Conference. pp. 108–122. Springer (2011)
6. Golbeck, J., Fragoso, G., Hartel, F., Hendler, J., Oberthaler, J., Parsia, B.: The national cancer institute's thesaurus and ontology. Journal of Web Semantics First Look 1_1_4 (2003)
7. Harrow, I., Jiménez-Ruiz, E., Splendiani, A., Romacker, M., Woollard, P., Markel, S., Alam-Faruque, Y., Koch, M., Malone, J., Waaler, A.: Matching disease and phenotype ontologies in the ontology alignment evaluation initiative. J Biomed Semantics **8**(1), 55 (2017)
8. Hayamizu, T.F., Mangan, M., Corradi, J.P., Kadin, J.A., Ringwald, M.: The adult mouse anatomical dictionary: a tool for annotating and integrating data. Genome biology **6**(3), 1–8 (2005)
9. Jiménez-Ruiz, E., Grau, B.C., Horrocks, I.: Exploiting the umls metathesaurus in the ontology alignment evaluation initiative.
10. Li, H., Dragisic, Z., Faria, D., Ivanova, V., Jiménez-Ruiz, E., Lambrix, P., Pesquita, C.: User validation in ontology alignment: functional assessment and impact. The Knowledge Engineering Review **34** (2019)
11. Rosse, C., Mejino Jr, J.L.: A reference ontology for biomedical informatics: the foundational model of anatomy. Journal of biomedical informatics **36**(6), 478–500 (2003)
12. Tomek, I.: Two modifications of cnn. IEEE Transactions on Systems, Man, and Cybernetics **SMC-6**(11), 769–772 (1976)
13. Zamazal, O., Svátek, V.: The ten-year ontofarm and its fertilization within the onto-sphere. Journal of Web Semantics **43**, 46–53 (2017)