Results for Matcha and Matcha-DL in OAEI 2023

Daniel Faria[1], Marta Silva[2], Pedro Cotovio[2], Lucas Ferraz[2], Laura Balbi[2] and Catia Pesquita[2]

[1]*INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal*
[2]*LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal*

Abstract

Matcha is an ontology matching system under development, designed to tackle long-standing challenges such as complex and holistic ontology matching. It incorporates all of the key algorithms from AgreementMakerLight (AML) over a novel broader core architecture, and includes several new algorithms. Matcha-DL augments Matcha to semi-supervised tasks, it uses trainable model to select and rank candidates proposed by Matcha. Matcha performed well overall, achieving the highest F-measure in 6 of the 18 distinct OAEI tasks and ranking in the top three in 9 others. Matcha-DL achieved the highest F-measure in 4 of the 5 semi-supervised BioML tasks.

1. Presentation of the System

1.1. State, Purpose, General Statement

Matcha is a novel ontology matching system that aims to tackle some long-standing challenges in the ontology matching field, namely: holistic ontology matching [1], complex ontology matching [2], and machine-learning based matching. Matcha incorporates all key algorithms from AgreementMakerLight (AML) on top of a novel core framework that encompasses matching of multiple ontologies and complex ontology matching by design. Its architecture is also more modular and extensible than AML's.

Matcha-DL is an augmentation of Matcha for semi-supervised learning that employs scores produced by several of Matcha's matching algorithms as input to train a dense neural network to discriminate and rank between a set of candidates produced by Matcha.

1.2. Specific Techniques Used

AML's lexical, and structural matching algorithms [3], as well as its background knowledge strategy [4] are all included in Matcha. There is also a revised implementation of the pattern-based complex matching algorithms that were present in AMLC [5]. As it is a system under development, some functionalities have yet to be implemented, such as alignment repair or interactive matching. There were also two implemented functionalities, namely complex matching and translation, that due to technical issues could not be used at the OAEI.

OM 2023: The 18th International Workshop on Ontology Matching collocated with the 22nd International Semantic Web Conference ISWC-2023 November 7th, 2023, Athens, Greece

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

A matching algorithm that uses Large Language Models (LLM) has been added to Matcha. The strategy relies on the conversion of the entities' labels and synonyms into embeddings, followed by a computation of the cosine similarity between embeddings. The embeddings can be obtained from any LLM model, although for this OAEI edition, we used the sentence-BERT [6] all-MiniLM-L6-v2 model[1] (the pretrained model with no fine-tuning).

Matcha's novel translation module is a neural-based many-to-many multilingual translation method that uses an Encoder-Decoder Long Short-Term Memory architecture that consists of two complex recurrent neural networks that act as an encoder and decoder pair. This mechanism solves a sequence-to-sequence prediction problem between two given source and target ontologies' languages.

The encoder maps the source language's label to a vector representation that serves as input to the decoder, that then maps the vector back to a translation of the label in the target language. The label translations are then added to the lexicons of the original ontologies.

For Matcha-DL, a specific pipeline was developed that incorporates nearly all of Matcha's matching algorithms to create an input for a dense neural network. Matcha-DL augments Matcha by learning to rank candidates produced by Matcha based on this input scores.

Matcha's matching algorithms are described in Table 1.

1.3. Adaptations Made for the Evaluation

The MELT [7] web-based package was implemented in Matcha for the required evaluation in OAEI.

Matcha-DL was mainly designed as a ranking model for local matching. It was however adapted to for global matching tasks by considering candidates produced by the Matcha core algorithm.

1.4. Link to the System and Parameters File

As Matcha is still under development, it is not publicly available. A public release is planned once the core development is completed, which is expected to be soon.

2. Results

Matcha's results for OAEI 2023 are summarized in Table 2, with the exception of the results for the BioML track, which are presented in Table 3, for both Matcha and Matcha-DL.

Matcha had good general performance, achieving the highest F-measure out of all systems in 8 of the 18 distinct OAEI tasks, while ranking in the top 3 in 9 others.

The participation in the complex track was hindered by the change in the definition of instances. In this year's datasets the entities shared a local name with different prefixes, but Matcha's algorithms rely on the entities being semantically equivalent – either by having the same URI or by being declared as *owl:sameAs.*

---

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Table 1
Summary of Matcha's key matching algorithms

Class Matching	
Instance-based Class Matcher	Matches classes based on overlapping individuals that instantiate them, computed through conservative instance matching algorithms
Lexical Matcher	Matches ontologies by finding literal full name matches between their lexicons. Weighs matches according to the provenance of the names
LLM Matcher	Matches ontologies by computing the cosine similarity between the embeddings of their lexicons
Mediating XRef Matcher	Matches ontologies by using cross-references and/or exact lexical matches between them and a third mediating ontology
String Matcher	Matches ontologies by measuring the maximum string similarity, using one of the four available string similarity measures
Word Matcher	Matches ontologies by measuring the word similarity, using a weighted Jaccard index
Instance Matching	
Attribute Matcher	Matches individuals by finding literal matches between the values of their annotation and data properties
Attribute String Matcher	Maps individuals by comparing their values through the ISub string similarity metric
Attribute to Lexicon Matcher	Maps individuals by comparing the lexicon entries of one with the values of the other using a combination of string and word matching algorithms

In regards to the Multifarm track, we had a character encoding issue that we were unable to resolve in time. In the Knowledge Graph track, Matcha only found matches for instances which was an unexpected result and requires further investigation.

In the Material Sciences and Engineering, some of the matches were between classes and object properties, which was puzzling considering that Matcha separates entities by type for the matching tasks, meaning there should be no mappings between different entity types. We suspect it could possibly be an issue with the ontologies' encoding.

Matcha-DL, through the incorporation of a relatively simple training procedure to augment Matcha, demonstrates surprising results on the semi-supervised Bio-ML tasks. Remarkably, it achieves the highest F-score in four out of five tasks. Comparative analysis with the Matcha algorithm underscores the discernible advantage of incorporating training within the context of ontology alignment.

3. Conclusions

There is a general improvement in Matcha's results when compared to the earlier version presented in OAEI 2022, although some further refinements are still required. Matcha achieved the highest F-measure in 6 of the 18 distinct OAEI tasks and ranked in the top three in 9 others.

Acknowledgements

This work was supported by FCT through the LASIGE Research Unit (UIDB/00408 /2020 and UIDP/00408/2020). It was also partially supported by the KATY project which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101017453. This work was supported in part by projet 41, HfPT: Health from Portugal, funded by the Portuguese Plano de Recuperação e Resiliência. Marta Silva was partially funded by FCT through the fellowship 2022.11895.BD.

References

[1] I. Megdiche, O. Teste, C. Trojahn, An extensible linear approach for holistic ontology matching, in: International Semantic Web Conference, Springer, 2016, pp. 393–410.
[2] É. Thiéblin, O. Haemmerlé, N. Hernandez, C. Trojahn, Survey on complex ontology matching, Semantic Web 11 (2020) 689–727. URL: https://doi.org/10.3233/SW-190366. doi:`10.3233/SW-190366`.
[3] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, F. M. Couto, The AgreementMakerLight Ontology Matching System, in: OTM Conferences - ODBASE, 2013, pp. 527–541.
[4] D. Faria, C. Pesquita, E. Santos, I. F. Cruz, F. M. Couto, Automatic Background Knowledge Selection for Matching Biomedical Ontologies, PLoS One 9 (2014) e111226.
[5] B. F. d. Lima, Breaking rules: taking Complex Ontology Alignment beyond rulebased approaches, Ph.D. thesis, 2021.
[6] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[7] S. Hertling, J. Portisch, H. Paulheim, MELT - matching evaluation toolkit, in: Semantic Systems. The Power of AI and Knowledge Graphs - 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9-12, 2019, Proceedings, 2019, pp. 231–245. URL: https://doi.org/10.1007/978-3-030-33220-4_17. doi:`10.1007/978-3-030-33220-4\_17`.

Table 2
Summary of Matcha’s OAEI 2023 results across 7 tracks

Task	Precision	Recall/ Coverage	F-measure	Run time (s)	Rank *
—— Anatomy ——					
Mouse-Human	0.951	0.931	0.941	54	1
—— Biodiversity & Ecology ——					
NCBITAXON-TAXREFLD Animalia	0.674	0.993	0.803	258	2
NCBITAXON-TAXREFLD Bacteria	0.58	0.994	0.743	14	3
NCBITAXON-TAXREFLD Chromista	0.624	0.984	0.764	48	3
NCBITAXON-TAXREFLD Fungi	0.785	0.998	0.879	103	2
NCBITAXON-TAXREFLD Plantae	0.741	0.989	0.847	196	3
NCBITAXON-TAXREFLD Protozoa	0.724	1.0	0.840	44	3
MACROALGAE-MACROZOOBENTHOS	0.2	0.5	0.285	7	4
FISH-ZOOPLANKTON	0.276	0.866	0.419	11	2
—— Common Knowledge Graphs ——					
Nell-DBpedia	0.99	0.87	0.93	113	2
Yago-Wikidata	0.99	0.90	0.94	436	1
—— Conference ——					
OntoFarm (rar2-M3)	0.62	0.62	0.62	-	4
—— Knowledge Graph ——					
Aggregate (overall)	0.55	0.84	0.66	51183	4
—— Material Sciences and Engineering ——					
First Test Case	0.273	0.130	0.176	28	1
Second Test Case	0.756	0.219	0.339	15	1
Third Test Case	0.949	0.889	0.918	23	1
—— Multifarm ——					
Different Ontologies	0.37	0.04	0.08	120	3
—— Food Nutritional Composition——					
Test Case Food V2	0.0656	0.3013	0.6567	53	1

* According to F-measure

Table 3
Summary of Matcha's and Matcha-DL's Bio-ML OAEI 2023 results.

Task	**Precision**	**Recall/ Coverage**	**F-measure**	**Rank** *	**MRR**	**Hits@1**
Semi-Supervised						
Matcha						
OMIM-ORDO	0.718	0.519	0.602	7	–	–
NCIT-DOID	0.839	0.750	0.792	8	–	–
SNOMED-FMA	0.846	0.502	0.630	8	–	–
SNOMED-NCIT (Pharm)	0.982	0.601	0.746	4	–	–
SNOMED-NCIT (Neoplas)	0.782	0.545	0.642	8	–	–
Matcha-DL						
OMIM-ORDO	0.745	0.732	0.738	1	0.811	0.780
NCIT-DOID	0.847	0.834	0.841	5	0.870	0.844
SNOMED-FMA	0.959	0.825	0.887	1	0.918	0.908
SNOMED-NCIT (Pharm)	0.903	0.872	0.888	1	0.931	0.917
SNOMED-NCIT (Neoplas)	0.806	0.714	0.757	1	0.829	0.806
Unsupervised						
Matcha						
OMIM-ORDO	0.781	0.509	0.617	6	-	-
NCIT-DOID	0.882	0.756	0.814	6	–	–
SNOMED-FMA	0.887	0.502	0.641	8	–	–
SNOMED-NCIT (Pharm)	0.987	0.607	0.752	2	–	–
SNOMED-NCIT (Neoplas)	0.838	0.551	0.665	5	–	–
Matcha-DL						
OMIM-ORDO	0.745	0.513	0.607	7	0.811	0.780
NCIT-DOID	0.847	0.586	0.693	10	0.870	0.844
SNOMED-FMA	0.960	0.602	0.740	4	0.918	0.908
SNOMED-NCIT (Pharm)	0.904	0.616	0.733	7	0.931	0.917
SNOMED-NCIT (Neoplas)	0.811	0.514	0.629	8	0.829	0.806

* According to F-measure