Results in OAEI 2024 for Matcha

Daniel Faria[1], Marta C. 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 designed to tackle long-standing challenges such as complex and holistic ontology matching. It incorporates all of the key algorithms from AgreementmakerLight over a novel broader core architecture that includes several new algorithms. In this year's edition, some strategies were modified to rectify some gaps found in last year, and a few new strategies were debuted, with particular note for the inclusion of Language Models in two of our algorithms. Matcha performed well overall, achieving the highest F-measure in 15 out of 43 distinct OAEI tasks and ranking in the top three in ten others.

1. Presentation of the System

1.1. State, Purpose, General Statement

Matcha is an ontology matching system that aims to tackle some long-standing challenges in the ontology matching field, namely complex ontology matching [1], holistic ontology matching [2], and machine-learning-based matching. Matcha builds upon the outstanding results of AgreementMakerLight (AML) [3] and incorporates its main algorithms combined with a core framework tailored to multi-ontology matching and complex matching and a number of new matching algorithms that explore language models.

1.2. Specific Techniques Used

Matcha includes all of AML's lexical and structural matching algorithms [4], as well as some of its background knowledge strategy [5]. For this year's OAEI, some matching techniques were revised, and some were newly developed.

One of the new matching algorithms uses a Language Model (LM) in order to go beyond the information that is explicitly stated in the ontology and exploit the context that labels and synonyms can provide when represented through a language model. The matching algorithm uses the LM to represent the entities' labels and synonyms as embeddings, which are subsequently compared through cosine similarity. Similarly to last year, we used the pre-trained sentence-BERT [6] all-MiniLM-L6-v2 model[1] without fine-tuning.

OM-2024: The 19th International Workshop on Ontology Matching collocated with the 23rd International Semantic Web Conference (ISWC 2024), November 11th, Baltimore, USA

[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
Language Model Matcher	Matches ontologies by computing the cosine similarity between the language model 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
Multilingual LM-based matcher	Maps individuals by comparing sentence representations of the source and target labels, obtained with a LM trained in a multilingual setting

Additionally, for any task that requires translation, we constructed a new translation module that uses a pre-trained multilingual translation LM, the "M2M100" [7], with 1.2B parameters and trained on over 100 languages. The model 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. The matching algorithm uses the encoder to map each of the source and target ontologies' labels to an embedding representation, followed by a computation of the cosine similarity between the embeddings to generate a mapping score.

Matcha's matching algorithms are described in Table 1.

1.3. Adaptations Made for the Evaluation

The MELT [8] web-based package was implemented in Matcha for the required evaluation in OAEI. Given two ontologies and a set of parameters, Matcha will generate a complete alignment between them according to the type of entities to be matched. For local alignment tasks, where each entity in the test set has a predetermined list of candidate matches, Matcha calculates scores for each candidate. These candidates are then ranked based on the highest score obtained from the various matching algorithms.

Matcha was packaged in a docker container for ease of sharing and running the evaluation, which included, for example, the files necessary for some of the algorithms, such as background knowledge ontologies used in some tracks.

2. Results

Matcha's results for OAEI are summarized in Table 2, with the exception of the results for the BioML track, which are presented in Table 3. Matcha performed well overall, achieving the highest F-measure out of all systems in 15 out of the 43 distinct OAEI tasks, while ranking in the top 3 in ten others.

2.1. Anatomy track

Matcha continues to excel in this track, placing first among all systems and with all evaluation metrics above a 0.9 (0.951 for precision, 0.931 for recall, 0.941 for F-measure). While not ranking first in precision, both precision and recall are very high, resulting in a high F-measure. It is interesting to note that the second-best system yields a 0.903 in F-measure.

2.2. Archaeology Multilingual track

This task had two participants: Matcha and LogMap with its three variants. Matcha achieved first place in F-measure in three out of ten tasks, but in some tasks, both systems achieved low-performance scores, including one task where all systems failed to return results. The results are heterogeneous, with some tasks achieving a high precision (including perfect precision for the de-de task), while others achieve values close to zero (six out of ten). In terms of recall, the results are also fairly heterogeneous, with the values varying between close to zero and 0.75.

With this being the first year that Matcha debuts the integrated MLLM-based translation module, we count on exploring other multilingual pretrained models and possibly performing a statistical analysis to understand the differences in language coverage and depth to sustain our future choice of a multilingual model for Matcha's translation module.

2.3. Biodiversity and Ecology track

This task also had two participants: Matcha and LogMap with its three variants. Matcha achieved first place in the F-measure in three of the nine tasks, while in most of the other tasks, it achieved scores that were very close to those obtained by LogMap. It is interesting to note that, in the NCBITAXON-TAXREFLD group, Matcha achieves perfect recall in two tasks, while in four others the value is very close to 1.0 (the lowest being 0.984). Precision is mostly consistent, oscillating between 0.57 and 0.74, with results being poorer in the MACROALGAE-MACROZOOBENTHOS and FISH-ZOOPLANKTON tasks, which are around 0.2.

2.4. Circular Economy track

In this new track, Matcha placed first out of all systems by F-measure. The results are moderate and close to other competing systems, with 0.393 precision and 0.611 recall. According to the organizers and an additional assessment performed, Matcha's optimal threshold could be set to 0.9, which would capture most true positives. On a less positive note, in terms of false positives and using a manual evaluation by the organizers, Matcha finds a fair amount of mappings that are probably due to the usage of either the same name or the same words, an interesting insight that could be used to further improve our strategies.

2.5. Conference track

Matcha tied in first place with another competing system, improving over last year's placement in all measures (precision, recall, and F-measure).

An additional evaluation was run to assess any differences in results from sharp, discrete, and continuous settings. From this assessment, it is noted that Matcha performs well in the sharp evaluation in terms of recall (0.67), but in the discrete uncertain setting, while its precision drops, recall improves to 0.77, indicating that it is successful at identifying uncertain matches. Matcha also appears to adapt well to the uncertain framework in the continuous setting, as its recall and F-measure remain relatively high at 0.75 and 0.71.

Regarding the evaluation performed based on logical reasoning, Matcha has 86 conservativity principle violations and 72 consistency principle violations in an alignment of 21 mappings. However, as the organizers note, conservativity principle violations can simply be false positives.

2.6. Digital Humanities track

Matcha achieves overall good results in this track, even if somewhat heterogeneous. Matcha ranks first in four out of the eight tracks, and in the top 3 in one other. Precision is high in some tracks (with a value of 1.0 in one of them), however in some others the value is close to zero, with Matcha yielding no results in one of these tasks. Recall suffers from less of this variability, with only the failed task having a value close to zero, and with good values for all others.

Similarly to the Archaeology Multilingual track, this track uses the MLLM-based translation module, which will be further explored and reviewed.

2.7. Food Nutritional Composition track

Matcha only competed in the "equal" relation testcase placing first against competing systems, however with a value of 0.1016 in F-measure which is lower than other tracks where it also places first. While Matcha is less precise than other systems (0.0611 against 0.1333), it compensates with its ten times higher recall (0.3013 against 0.0274). This track poses challenges that current systems are clearly not well equipped to handle.

2.8. Knowledge Graph track

Matcha places last in this track when assessing the aggregated results. Looking at class mappings, Matcha has a high performance overall with 0.97 of precision, 0.8 of recall, and 0.87 of F-measure, outperforming both competing systems and the baselines. All systems fail at finding property mappings, and as for instance mappings, Matcha has a lower performance with 0.55 of precision, 0.86 of recall and 0.63 of F-measure, finding far more mappings than other systems (249510 mappings versus 6653.8 by the next system) which decreases precision significantly.

When looking at each of the test cases, a pattern emerges where Matcha has lower precision and higher recall when compared to all other systems. However the precision values are low enough that cannot be compensated by the high recall, and therefore lead Matcha to place last in four of the five test cases, placing second in the remaining one, according to F-measure.

In this track, two main problems arise which need to be assessed and corrected: the lack of property mappings and the excessive amount of instance mappings produced, which directly influence the system's precision.

2.9. Multifarm track

Matcha's performance in this track is fairly balanced considering a new strategy of using LLMs for multilingual machine-translation was debuted this year.

This year Matcha ranked second out of four systems, outperforming last year's scores where Matcha competed without the LLM module and ranked 4th out of four, with a clear improvement in recall and F-measure. Although Matcha's running time is within the same order of magnitude as other competing systems, we recognize that it is very time-consuming in its current iteration and could be optimized in future versions of the system.

2.10. Bio-ML track

This year marks Matcha's first time competing in the local alignment challenges of this track. While Matcha's Bio-ML rankings based on F-score were moderate compared to the other participating models, Matcha demonstrated a stronger relative performance when considering MRR, especially in the unsupervised setting. Matcha's middle-ranking F-scores were caused by largely high precisions paired with relatively low recalls, a trend also evident among most other participating systems, highlighting that the challenge of improving recall without compromising precision still remains an open issue. Overall Matcha results to note in the Bio-ML track include: a top-3 MRR ranking placement in 3 of the 5 tasks in the unsupervised setting; a first and second place in MRR ranking in the unsupervised and supervised settings of the SNOMED-FMA (body) task, respectively; and a second place in the F-score ranking in the unsupervised SNOMED-NCIT (pharm) task.

3. Conclusions

Matcha achieved the highest F-measure in 15 out of the 43 distinct OAEI tasks and ranked in the top 3 in ten others, making it overall the second-best system that competed this year.

This year a new approach for the translation model was debuted that allowed Matcha to improve its rank in the multifarm track, and place fairly well in the new tracks of archaeology-multilingual and digital humanities. Moreover, across all tasks, Matcha tends to outperform other systems in recall, while tending to underperform in precision, sometimes due to an exaggerated number of mappings found that turn out to be false positives. Some tracks require further review, such as the knowledge graph track where Matcha fails to find any property mappings.

Acknowledgements

This work was supported by FCT through fellowships 2022.11895.BD (Marta Silva), 2022.10557.BD (Pedro Cotovio), and through the KATY project through fellowships R881.7 (Laura Balbi), and the LASIGE Research Unit, ref. UIDB/00408/2020 (https://doi.org/10.54499/UIDB/00408/2020) and ref. UIDP/00408/2020 (https://doi.org/10.54499/UIDP/00408/2020). It was 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, and it was also partially supported by project 41, HfPT: Health from Portugal, funded by the Portuguese Plano de Recuperação e Resiliência.

References

[1] 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.

[2] I. Megdiche, O. Teste, C. Trojahn, An extensible linear approach for holistic ontology matching, in: International Semantic Web Conference, Springer, 2016, pp. 393–410.

[3] D. Faria, E. Santos, B. S. Balasubramani, M. C. Silva, F. M. Couto, C. Pesquita, Agreementmakerlight, Semantic Web (2023) 1–13.

[4] 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.

[5] 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.

[6] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[7] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, A. Joulin, Beyond english-centric multilingual machine translation, 2020. arXiv:2010.11125.

[8] 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 2024 results across 7 tracks

Task	Precision	Recall/ Coverage	F-measure	Run time (s)	Rank *
		—— Anatomy ——			
Mouse-Human	0.951	0.931	0.941	42	1
		—— Archaelogy Multilingual ——			
idai-pactols_de-de	1.0	0.12	0.21	-	4
idai-pactols_de-en	0.02	0.06	0.03	-	4
idai-pactols_de-fr	0.04	0.18	0.07	-	4
idai-pactols_de-it	0.00	0.00	0.00	-	4
idai-pactols_en-en	0.75	0.50	0.60	-	1
idai-pactols_en-fr	0.03	0.33	0.05	-	4
idai-pactols_en-it	0.00	0.00	0.00	-	4
idai-pactols_fr-fr	0.25	0.25	0.25	-	1
idai-pactols_fr-it	0.00	0.00	0.00	-	-
idai-pactols_it-it	0.30	0.75	0.43	-	1
		—— Biodiversity & Ecology ——			
ENVO-SWEET	0.758	0.808	0.783	40	1
NCBITAXON-TAXREFLD Animalia	0.675	0.994	0.804	78	1
NCBITAXON-TAXREFLD Bacteria	0.578	1.0	0.732	4	2
NCBITAXON-TAXREFLD Chromista	0.623	0.984	0.763	14	4
NCBITAXON-TAXREFLD Fungi	0.785	0.998	0.879	36	1
NCBITAXON-TAXREFLD Plantae	0.741	0.993	0.849	61	2
NCBITAXON-TAXREFLD Protozoa	0.723	1.0	0.839	11	2
MACROALGAE-MACROZOOBENTHOS	0.2	0.5	0.286	5	4
FISH-ZOOPLANKTON	0.277	0.867	0.419	8	2
		—— Circular Economics ——			
CEON-BiOnto	0.265	0.722	0.388	-	3
		—— Conference ——			
OntoFarm (rar2-M3)	0.66	0.63	0.64	-	1
		—— Digital Humanities ——			
arch1_defc-pactols	1.0	0.9	0.95	-	1
arch2_idai-pactols	0.31	0.24	0.27	-	4
arch3_ironagedanube-pactols	0.67	0.8	0.73	-	1
arch4_pactols-parthenos	0.83	0.83	0.83	-	1
cult1_idai-parthenos	0.0	0.0	0.0	-	4
cult2_oeai-parthenos	0.9	0.74	0.81	-	1
dhcs1_dha-unesco	0.08	0.6	0.14	-	4
dhcs2_tadirah-unesco	0.48	0.67	0.56	-	2
		—— Food Nutritional Composition——			
Test Case Food V2	0.0611	0.3013	0.1016	47	1
		—— Knowledge Graph ——			
Aggregated (overall)	0.55	0.84	0.63	139696	5
		—— Multifarm ——			
Aggregated	0.21	0.44	0.28	18540	2

* According to F-measure

Table 3
Summary of Matcha's Bio-ML OAEI 2024 results.

Task	Precision	Recall/ Coverage	F-measure	MRR	Hits@1	Rank *
			Semi-Supervised			
OMIM-ORDO	0.718	0.519	0.602	0.815	0.782	3
NCIT-DOID	0.839	0.750	0.792	0.902	0.873	4
SNOMED-FMA	0.846	0.502	0.630	0.950	0.935	2
SNOMED-NCIT (Pharm)	0.982	0.601	0.746	0.936	0.921	4
SNOMED-NCIT (Neoplas)	0.782	0.545	0.642	0.899	0.936	4
			Unsupervised			
			Matcha			
OMIM-ORDO	0.781	0.509	0.617	0.815	0.782	3
NCIT-DOID	0.882	0.756	0.814	0.902	0.873	3
SNOMED-FMA	0.887	0.502	0.641	0.950	0.935	1
SNOMED-NCIT (Pharm)	0.987	0.607	0.752	0.936	0.921	4
SNOMED-NCIT (Neoplas)	0.838	0.551	0.665	0.899	0.936	4

* According to MRR