Matcha results in OAEI 2025

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

[1]LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2]INESC-ID, Instituto Superior Técnico, 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. Matcha performed well overall, achieving the highest F-measure in ten out of 32 distinct OAEI tasks and ranking in the top three in twelve others. In the complex track, Matcha achieved the highest F-measure in seven tasks using the Graph Edit Distance and two using the Tree Edit Distance.

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].

Novel algorithms include Language Models (LM) in their strategies, 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. In both the translation tasks and the complex track, we used the pre-trained sentence-BERT [6] all-MiniLM-L6-v2 model[1] without fine-tuning. The matching algorithm uses the LM to represent the entities' labels and synonyms as embeddings, which are subsequently compared through cosine similarity. To find combinations of classes the recursive strategy described in [7] is used.

In the Bio-ML track for the ranking task, the scores for all previously described matchers are computed and a maximum function is applied to determine the final score for each candidate mapping.

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

OM 2025: The 20th International Workshop on Ontology Matching collocated with the 24th International Semantic Web Conference (ISWC 2025), November 2nd, 2025, Nara, Japan

[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

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 and the scripts necessary to run the language models.

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, and the results for the complex track, which are presented in Table 4. Matcha achieved the highest F-measure in ten out of 32 distinct OAEI tasks and ranking in the top three in twelve others. In the complex track, Matcha achieved the highest F-measure in seven tasks using the Graph Edit Distance and two using the Tree Edit Distance.

2.1. Anatomy track

With the same results as last year, Matcha placed first among all systems achieving scores higher than 0.9 in all evaluation metrics (0.951 for precision, 0.931 for recall, 0.941 for F-measure).

2.2. Archaeology Multilingual track

There was an increase in participating systems from two last year to four this year (LogMap competes with three variants). Matcha placed first in three out of ten tasks and failing to produce an alignment in

one of them (where only one system had any results), similarly to last year. The results are again very heterogeneous, with precision raging from 0.17 to 1.0 (in the idai-pactols_de-de task) and recall from 0.06 to 0.75. Unsurprisingly the best results are mostly achieved in the single-language tasks.

Due to some technical issues, the language model was changed from last year's MLLM to sentence-BERT, and it is interesting to note that the overall pattern of results were maintained despite this change. One of the improvement points to consider for the next edition is the runtime, as Matcha takes close to 11 minutes while most systems take under a minute to achieve better or comparable results.

2.3. Circular Economy track

This year's results in this track were lower than the previous year requiring some further exploration, with Matcha placing last behind AgentOM and two LogMap variants. As mentioned by the track's organizers, Matcha finds a large number of false positives, mostly in situtations where entities share the same name, which greatly impacts precision (0.158). Recall scores far higher with a 0.917, being the second best value among all systems.

2.4. Complex track

Matcha's complex matcher finds simultaneously simple equivalences, subsumption relationships, and 1:n complex equivalences. For the unpopulated conference task only complex mappings were evaluated and Matcha was unable to find any true positives.

Three new evaluations were debuted this year that evaluate mappings in the alignment based on their content and structure. Matcha performed much better in the metrics that took into account simple mappings (class evaluation and GED evaluation), since it actually outputted few complex mappings. Further exploration is necessary in this track in order to produce more (and correct) complex mappings.

Matcha competed as the only system in the new biomedical dataset achieving results that serve as a good starting point between 0.465 and 0.554 in F-measure.

2.5. Conference track

Compared to last year, Matcha improved on precision but lowered recall resulting in an F-measure that is 0.01 lower (0.63), placing third. Matcha outperforms both baselines.

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 precision (0.85), but in the discrete uncertain setting, while its precision drops, recall improves to 0.77 and then remains at a 0.75 in the continuous evaluation, indicating that it is successful at identifying uncertain matches.

Regarding the evaluation performed based on logical reasoning, Matcha has 90 conservativity principle violations and 115 consistency principle violations in an alignment of 21 mappings. which is an increase from last year and requires further exploration. As the organizers note, conservativity principle violations can be false positives.

2.6. Digital Humanities track

Matcha improved over last year's results placing first in five tasks out of eight tasks. Futhermore, overall it achieves the best average F1-score of 0.64 and the task that failed to produce an alignment in the previous edition was successfully completed this year. The results are somewhat heterogeneous with precision raging from 0.31 to 1.0 (in the arch1_defc-pactols) and recall from 0.23 to 1.0 (in the arch2_idai-pactols). As in the previous year, recall is less variable with more values falling on the top end of the scale (six out of eight are higher than 0.7).

Similarly to the Archaeology Multilingual track, this track uses sentenceBERT as the new language model for the translation module, which still requires further exploration and review.

2.7. Knowledge Graph track

Matcha places second to last overall, failing to produce an alignment for one of the tasks and producing no property mappings. By not considering empty, erroneous, or not generated alignments, Matcha improves on its results, with precision increasing from 0.54 to 0.68, recall from 0.71 to 0.88, and consequently F-measure from 0.61 to 0.76.

In aligning instances, Matcha finds far more mappings than other systems (29113 mappings when the next system finds 6653) which will significantly decrease precision.

When looking at each of the test cases, a pattern emerges where Matcha has lower precision and higher recall when comparing all systems. However the precision values are low enough that they cannot be compensated by the high recall, leading Matcha to rank very low.

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.

2.8. Multifarm track

Matcha placed second overall by F-measure in this track. Both precision and recall are low and almost equal at 0.26 and 0.25, respectively. Despite other systems having a higher precision, lower recall values decrease the F-measure significantly. The runtime is still a necessary point of improvement as it takes significantly longer than other systems.

2.9. Bio-ML track

This year, Matcha maintained consistent performance without any major system updates, which is a positive outcome given the participation of four new machine learning–based systems. While Matcha's Bio-ML rankings based on F-score were moderate compared to the other participating models, it demonstrated a stronger relative performance in terms of Mean Reciprocal Rank (MRR), particularly in the unsupervised setting. The middle-range F-scores mainly reflected high precision paired with comparatively low recall, a pattern also observed among most other participating systems, highlighting that the challenge of improving recall without compromising precision remains an open issue. Overall, key Matcha results in the Bio-ML track include: a top-3 MRR ranking in four of the five tasks in the unsupervised setting; first and second places in MRR ranking in the unsupervised and supervised settings of the SNOMED–FMA task, respectively; and a second place in the SNOMED–NCIT (pharm) task.

3. Conclusions

Matcha achieved the highest F-measure in ten out of the 32 distinct OAEI tasks and ranked in the top three in twelve others. Overall, results stayed mostly the same compared to the previous edition. It is interesting to note that changes in the language model of the translation module had limited impact in the archaeology-multilingual and digital humanities tracks. The new results in the complex track are interesting to note, as they combine different types of mappings and were evaluated following new strategies. Some issues failed to be resolved from last year, such as the the ones linked to the knowledge graph track, meaning Matcha still requires some further review.

Acknowledgements

This work was supported by FCT through fellowships https://doi.org/10.54499/2022.11895.BD (Marta Silva), https://doi.org/10.54499/2022.10557.BD (Pedro Cotovio), https://doi.org/10.54499/2025.04034.BD (Lucas Ferraz) and https://doi.org/10.54499/2024.01208.BD (Laura Balbi), and through LASIGE Research Unit, ref. UID/408/2025. 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, by the CancerScan project by the EU's HORIZON Europe research and innovation programme under grant agreement No 101186829, and by project 41, HfPT: Health from Portugal, funded by the Portuguese Plano de Recuperação e Resiliência.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

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] M. C. Silva, D. Faria, C. Pesquita, Complex multi-ontology alignment through geometric operations on language embeddings., in: ECAI, 2024, pp. 1333–1340.

[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 2025 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	0.17	0.06	0.09	-	5
idai-pactols_de-fr	0.33	0.12	0.17	-	5
idai-pactols_de-it	0.40	0.12	0.18	-	4
idai-pactols_en-en	0.75	0.50	0.60	-	1
idai-pactols_en-fr	0.33	0.17	0.22	-	4
idai-pactols_en-it	0.50	0.17	0.25	-	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
—— Circular Economics ——					
CEON-BiOnto	0.632	0.828	0.716	-	2
CEON-MATONTO	0.319	0.938	0.476	-	3
—— Conference ——					
OntoFarm (rar2-M3)	0.77	0.53	0.63	-	3
—— Digital Humanities ——					
arch1_defc-pactols	1.0	0.9	0.95	-	1
arch2_idai-pactols	0.45	1.0	0.63	-	1
arch3_ironagedanube-pactols	0.31	0.24	0.27	-	5
arch4_pactols-parthenos	0.80	0.23	0.36	-	5
cult1_idai-parthenos	0.67	0.80	0.73	-	1
cult2_oeai-parthenos	0.9	0.74	0.81	-	1
dhcs1_dha-unesco	0.83	0.83	0.83	-	1
dhcs2_tadirah-unesco	0.36	0.93	0.52	-	2
—— Knowledge Graph ——					
Aggregated (overall and non-empty)	0.68	0.88	0.76	6205	7
—— Multifarm ——					
Aggregated	0.26	0.25	0.26	24480	2

*According to F-measure

Table 3
Summary of Matcha's Bio-ML OAEI 2025 results.

Task	Precision	Recall/ Coverage	F-measure	MRR	Hits@1	Rank *
Semi-Supervised						
NCIT-DOID	0.839	0.750	0.792	0.902	0.873	3
OMIM-ORDO	0.718	0.519	0.602	0.815	0.782	3
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	2
SNOMED-NCIT (Neoplas)	0.782	0.545	0.642	0.899	0.936	4
Unsupervised						
Matcha						
NCIT-DOID	0.882	0.756	0.814	0.902	0.873	3
OMIM-ORDO	0.781	0.509	0.617	0.815	0.782	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	2
SNOMED-NCIT (Neoplas)	0.838	0.551	0.665	0.899	0.936	4

* According to MRR

Table 4
Summary of Matcha's OAEI 2025 results for the class evaluation of the complex track

Task	Precision	Recall/ Coverage	F-measure	Run time (s)	Rank *
—— Complex (GED evaluation) ——					
cmt-conference	0.526	0.365	0.431	-	1
cmt-ekaw	0.529	0.327	0.404	-	1
conference-ekaw	0.587	0.595	0.591	-	1
cree-swo	0.500	0.064	0.113	-	1
hydro3-swo	0.645	0.507	0.568	-	1
hydrOntology_native-swo	0.372	0.062	0.106	-	-
hydrontology_translated-swo	0.421	0.303	0.353	-	1
enslaved-wikidata	0.048	0.688	0.091	-	1
gbo-gmo	0.287	0.467	0.355	-	2
popgbo-popgmo	0.245	0.391	0.301	-	2
hp	0.465	0.465	0.465	-	-
mp	0.529	0.529	0.529	-	-
wbp	0.554	0.554	0.554	-	-
—— Complex (TED evaluation) ——					
cmt-conference	0.000	0.000	0.000	-	-
cmt-ekaw	0.000	0.000	0.000	-	3
conference-ekaw	0.000	0.000	0.000	-	-
cree-swo	0.077	0.011	0.019	-	2
hydro3-swo	0.000	0.000	0.000	-	-
hydrOntology_native-swo	0.079	0.004	0.008	-	1
hydrontology_translated-swo	0.569	0.202	0.299	-	1
enslaved-wikidata	0.001	0.002	0.001	-	2
gbo-gmo	0.003	0.002	0.002	-	2
popgbo-popgmo	0.002	0.004	0.003	-	2

Task	Correct	Contains reference	Contained in reference	Overlap	Incorrect
—— Complex (Class evaluation) ——					
cmt-conference	8	0	2	1	15
cmt-ekaw	8	0	2	1	12
conference-ekaw	14	0	1	0	20
cree-swo	3	0	0	0	1
hydro3-swo	13	0	1	0	6
hydrOntology_native-swo	2	0	0	0	6
hydrontology_translated-swo	13	0	0	5	35
enslaved-wikidata	11	0	0	0	112
gbo-gmo	17	0	0	2	13
popgbo-popgmo	11	6	0	0	167
hp	826	156	393	2366	2246
mp	2441	689	543	3131	3318
wbp	56	101	14	401	390