LSMatch and LSMatch-Multilingual Results for OAEI 2023

Abhisek Sharma[1,*], Sarika Jain[1]

[1]National Institute of Technology Kurukshetra, India

Abstract

The Large-Scale Ontology Matching System (LSMatch and LSMatch-Multilingual) and its findings using OAEI 2023 datasets are presented in this paper. A string similarity and synonyms matcher is used in the element-level and label-based ontology matching system called LSMatch. Same configuration in addition with MyMemory translation memory is used in the creation of multilingual capable system called LSMatch-Multilingual. The system(s) is/are capable of identifying classes, instances, and properties (both in monolingual and multilingual settings) between two ontologies. This year LSMatch and LSMatch-Multilingual are collectively participating on OAEI's five tracks—Anatomy, Conference, Multifarm, Common Knowledge Graphs, and Knowledge Graph. LSMatch has shown encouraging outcomes across all five tracks.

Keywords

Ontology Matching, Knowledge Schema, Alignment, String similarity, Synonym matcher.

1. Presentation of the system

1.1. State, purpose, general statement

LSMatch (Large Scale Ontology Matching System) is an ontology matching system that finds correspondences between ontologies using lexical properties. It employs the Levenshtein string similarity measure and the synonyms matcher, which employs background knowledge containing synonyms to filter out concepts with similar meanings but different lexical representations [1]. For multilingual LSMatch uses MyMemory translation memory. This is LSMatch's third OAEI appearance, and it was tested on five tracks: Anatomy, Conference, Multifarm, Common Knowledge Graphs, and Knowledge Graph. The LSMatch system was wrapped in the MELT framework [2], and it is performing at par with other systems, in Multifarm LSMatch-Multilingual got highest F1-score.

1.2. Specific techniques used

The current version of LSMatch (as compared to last year's submission) is now capable addresses both monolingual and multilingual ontology alignments. The working of the LSMatch system

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

*Corresponding author.

abhisek_61900048@nitkkr.ac.in (A. Sharma); jasarika@nitkkr.ac.in (S. Jain)

is shown in figure 1. We introduce the multiple parts of the system by taking two Knowledge schemas/ontologies. LSMatch system takes input in any format and loads the input schemas/ontologies as RDF graphs. After extracting classes, properties, and instances we perform stemming, removing stopwords and non-alphabetic characters, and normalizing letters. Then we pass the ontology concepts from Levenshtein and synonyms matcher modules. The underline modules have following functionality:

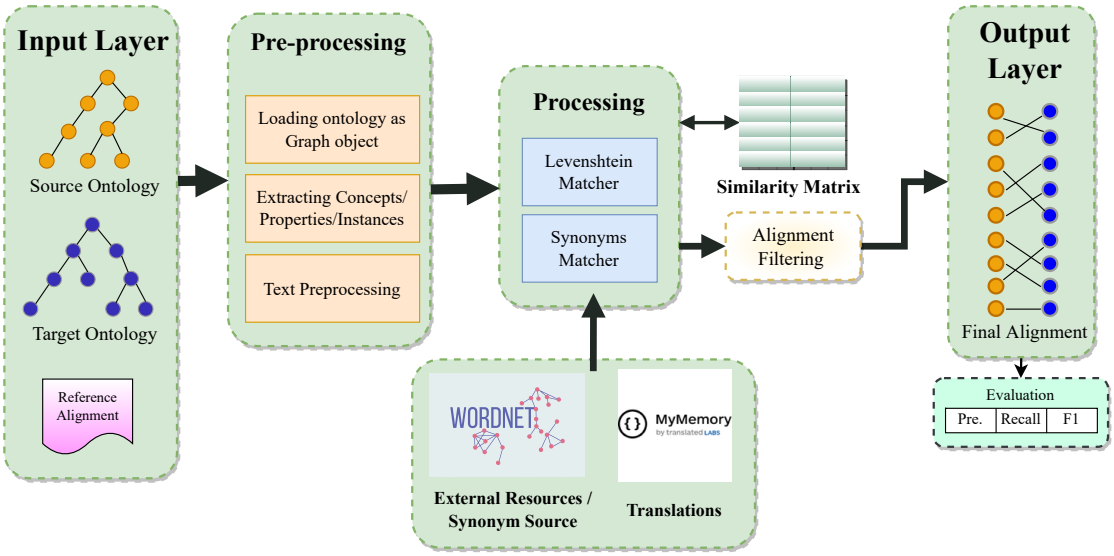

Figure 1: Combined architecture of LSMatch and LSMatch-Multilingual systems

- Levenshtein matcher: The LSMatch uses a string similarity matcher that calculates Levenshtein distance between the concepts [3]. The concepts are represented as rdfs:label or directly as the class name in the ontologies. The official definition of Levenshtein distance is stated as "The smallest number of insertions, deletions, and substitutions required to change one string or tree into another"[1].
- Background knowledge [4]: To identify different lexical representations, LSMatch uses a synonym matcher that fetches synonyms Wordnet [5]. Python's nltk library is used for wordnet inclusion.
- Synonym Matcher: LSMatch fetches synonyms from wordnet. Although we have pre-fetched the synonyms but during the execution, the concepts are cross-checked whether the synonyms for every concept are present or not. If some concept doesn't have synonyms pre-fetched for it, we fetch them on the fly.
- Translations[2]: for translations we have used MyMemory's translations memory as its provide good translations, is free, and is the world's largest Translation Memory.

For the purpose of storage and retrieval of alignments LSMatch uses dictionary. In the dictionary, we store information as <key, value> pairs where key is hashed [6, 7]. LSMatch

[1]https://xlinux.nist.gov/dads/HTML/Levenshtein.html
[2]https://mymemory.translated.net/

stores the alignments received from both the matchers along with the similarity score. We target storing and updating the scores of pairs multiple times during the alignment process and having hashed keys allow us to do that efficiently. By default, LSMatch keeps all the alignments with a combined score (Levenshtein + Synonym) of 0.5 or above to check the alignments over variable thresholds. For the final selection of alignments the current version of LSMatch has used 0.95 as the threshold.

2. Results

This section describes the results of the LSMatch and LSMatch-multilingual system collectively on five tracks namely: Anatomy, Conference, Multifarm, Common Knowledge Graphs, and Knowledge Graph. The results are presented collectively in Table 1. Differences from OAEI2022 [8] are discussed in the subsections below.

2.1. Anatomy

In anatomy overall result is same as last year with no change in performance results.

2.2. Conference

For conference track the result are exactly same as last year as due to some error we had to use the last year's LSMatch for this track, because of which the results are identical.

2.3. Multifarm

This is the second entry of LSMatch in Multifarm track. For this track we specifically developed LSMatch-multilingual. This year LSMatch-multilingual saw improvement in time, though the values of performance other than time were identical.

2.4. Bio-ML

The Bio-ML track is Machine Learning (ML) friendly Biomedical track. This track supersedes the previous largebio and phenotype tracks. There are 5 tasks in total (on which LSMatch was tested), all Equivalent matching have been performed with 5 ontology pairs, OMIN-ORDO(Disease), NCIT-DOID(Disease), SNOMED-FMA(Body), SNOMED-NCIT(Pharm), and SNOMED-NCIT(Neoplas). On OMIN-ORDO(Disease) and NCIT-DOID(Disease) LSMatch got average results. On SNOMED-FMA(Body), LSMatch has 6th best precision out of 9. On SNOMED-NCIT(Pharm) and SNOMED-NCIT(Neoplas), LSMatch has 2nd best precision just after LogMap-Lite. All the above stated resutls are on Unsupervised (90% Test Mapping). For Semi-supervised(70% Test Mappings), LSMatch was not tested this year on this track.

2.5. Common Knowledge Graphs

This year the performance of LSMatch on common Knowledge Graph track are identical in Nell-DBPedia task. Though there is a 0.01% improvement in Yago-Wikidata.

Table 1
Result summary of LSMatch and LSMatch-multilingual at OAEI 2023 and OAEI 2022

Task	Year	Precision	F1	Recall
––Anatomy––				
Mouse-Human	2023	0.952	0.761	0.634
Mouse-Human	2022	0.952	0.761	0.634
––Conference––				
OntoFarm (rar2-M3)	2023	0.83	0.55	0.41
OntoFarm (rar2-M3)	2022	0.83	0.55	0.41
OntoFarm (Sharp)	2023	0.88	0.57	0.42
OntoFarm (Sharp)	2022	0.88	0.57	0.42
OntoFarm (Discrete)	2023	0.88	0.66	0.53
OntoFarm (Discrete)	2022	0.87	0.66	0.53
OntoFarm (Continuous)	2023	0.88	0.67	0.54
OntoFarm (Continuous)	2022	0.88	0.67	0.54
––Bio-ML (Unsupervised (90% Test Mapping))––				
Equivalent Matching Results for OMIM-ORDO (Disease)	2022	0.65	0.329	0.221
Equivalent Matching Results for NCIT-DOID (Disease)	2022	0.719	0.633	0.565
Equivalent Matching Results for SNOMED-FMA (Body)	2022	0.809	0.132	0.072
Equivalent Matching Results for SNOMED-NCIT (Pharm)	2022	0.982	0.706	0.551
Equivalent Matching Results for SNOMED-NCIT (Neoplas)	2022	0.902	0.377	0.238
––Bio-ML (Semi-supervised (70% Test Mapping))––				
Equivalent Matching Results for OMIM-ORDO (Disease)	2022	0.594	0.325	0.223
Equivalent Matching Results for NCIT-DOID (Disease)	2022	0.665	0.611	0.565
Equivalent Matching Results for SNOMED-FMA (Body)	2022	0.762	0.128	0.07
Equivalent Matching Results for SNOMED-NCIT (Pharm)	2022	0.976	0.702	0.548
Equivalent Matching Results for SNOMED-NCIT (Neoplas)	2022	0.877	0.374	0.238
––Multifarm––				
Multifarm	2023	0.68	0.47	0.36
Multifarm	2022	0.68	0.47	0.36
––Common KG Track––				
Nell-DBPedia	2023	0.96	0.84	0.75
Nell-DBPedia	2022	0.96	0.84	0.75
Yago-Wikidata	2023	0.97	0.76	0.63
Yago-Wikidata	2022	0.96	0.76	0.63

––Knowledge Graph Track––

Year	Class			Property			Instance			Overall		
	P	F1	R	P	F1	R	P	F1	R	P	F1	R
2023	0.97	0.78	0.64	0.73	0.71	0.69	0.66	0.63	0.6	0.66	0.63	0.61
2022	0.97	0.78	0.64	0.73	0.71	0.69	0.66	0.63	0.6	0.66	0.63	0.61

2.6. Knowledge Graph

The performance of LSMatch is identical to last year's result in all aspects (Time, Precision, Recall, F1).

3. Conclusion

This year, the combincation of systems (LSMatch and LSMatch Multilingual) was collectively tested on five tracks, i.e., Anatomy, Conference, Multifarm, Common Knowledge Graphs, and Knowledge Graph. The system achieved considerably good precision in all the tracks but lacked behind in recall. In future versions, we will be adding a set of matchers and working to improve the utilization of background knowledge by which we can find better correlations between concepts that are not properly aligned using just the lexical measures.

References

[1] S. Zhang, Y. Hu, G. Bian, Research on string similarity algorithm based on levenshtein distance, in: 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), IEEE, 2017, pp. 2247–2251.

[2] S. Hertling, J. Portisch, H. Paulheim, Melt-matching evaluation toolkit, in: International conference on semantic systems, Springer, Cham, 2019, pp. 231–245.

[3] T. T. A. Nguyen, S. Conrad, Ontology matching using multiple similarity measures, in: 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), volume 1, IEEE, 2015, pp. 603–611.

[4] Z. Aleksovski, W. Ten Kate, F. Van Harmelen, Exploiting the structure of background knowledge used in ontology matching., in: Ontology Matching, 2006, p. 13.

[5] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (1995) 39–41.

[6] P. Ochieng, S. Kyanda, Large-scale ontology matching: State-of-the-art analysis, ACM Computing Surveys (CSUR) 51 (2018) 1–35.

[7] S. Anam, Y. S. Kim, B. H. Kang, Q. Liu, Review of ontology matching approaches and challenges, International Journal of Computer Science and Network Solutions 3 (2015) 1–27.

[8] A. Sharma, A. Patel, S. Jain, Lsmatch and lsmatch-multilingual results for oaei (2022).