

LSMatch Results for OAEI 2021

Abhisek Sharma¹, Archana Patel², and Sarika Jain¹

¹ Department of Computer Applications, National Institute of Technology
Kurukshetra, India

{abhisek_61900048, jasarika}@nitkkr.ac.in

² Department of Software Engineering, Eastern International University, Vietnam
archanamca92@gmail.com

Abstract. This paper presents the Large Scale Ontology Matching System (LSMatch) and its results on OAEI 2021 datasets. LSMatch is an element-level and label-based ontology matching system that uses string similarity and synonyms matcher. The current version of the system is focused on finding similarities between the classes of the two ontologies. This is the first participation of LSMatch in the OAEI campaign on five tracks, namely Anatomy, Conference, Disease and Phenotype, Common Knowledge Graphs, and Knowledge Graph. LSMatch has demonstrated promising results in all five tracks. We also discuss the strengths and weaknesses of the LSMatch system.

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 exploiting lexical properties to find correspondences between ontologies. It uses Levenshtein string similarity measure and synonyms matcher, which utilizes background knowledge containing synonyms to filter out concepts that are similar by meaning but have different lexical representations [1]. This is LSMatch's first OAEI participation, and it got tested on 5 tracks, i.e., Anatomy, Conference, Disease and Phenotype, Common Knowledge Graphs, and Knowledge Graph. LSMatch system was wrapped using the MELT framework [2], and it is performing at par with some of the other systems in tracks and achieving best precision in Anatomy and Conference track.

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1.2 Specific techniques used

The current version of LSMatch addresses monolingual ontology alignments, i.e., the concepts of the ontologies are in the same language, English [3]. We have called ontology as knowledge schema (KS) because the LSMatch system matches the classes only. The working of the LSMatch system is shown in figure 1. We introduce the multiple parts of the system by taking two Knowledge schemas (KS1 and KS2) as input to show the final set of alignments. LSMatch system takes input in any format and loaders loads the input KS (KS1 and KS2) as RDF graphs.

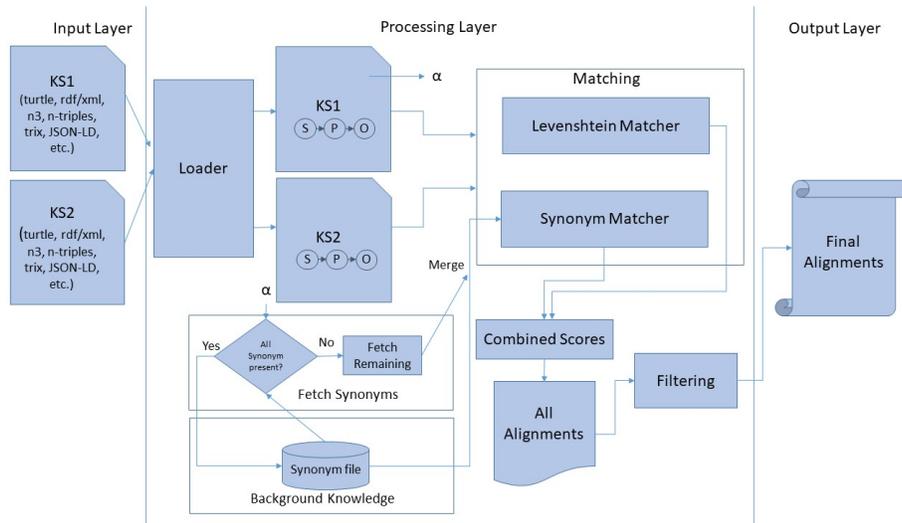


Fig. 1. Overview of LSMatch system

- **Levenshtein matcher:** The LSMatch uses a string similarity matcher that calculates Levenshtein distance between the concepts [4]. 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” [5].
- **Background knowledge [6]:** To identify different lexical representations, LSMatch uses a synonym matcher that fetches synonyms from `thesaurus.com` through their API [7]. For immediate availability of synonyms at the time of matching, we have pre-fetched the synonyms and kept them in JSON format.
- **Synonym Matcher:** LSMatch fetches synonyms from `thrsaurus.com`. 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.

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 [8, 9]. LSMatch 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 system on five tracks namely: Anatomy, Conference, Disease and Phenotype, Common Knowledge Graphs, and Knowledge Graph.

2.1 Anatomy

The Anatomy track consists of finding the alignments between the Adult Mouse Anatomy and the NCI Thesaurus describing the human anatomy. For the evaluation a 16GB machine was used¹. Table 1 shows the performance of LSMatch system on anatomy track. LSMatch generated 940 total correspondences, out of which 937 were true positives, and 3 were false positive. In Anatomy is achieving best precision along side the baseline matcher.

2.2 Conference

The Conference track contains 16 ontologies from the same domain (conference organization). Seven ontologies are involved in the reference alignment: Cmt, ConfTool, Edas, Ekaw, Iasted, Sigkdd, Sofsem. Table 2 shows the performance of LSMatch system on conference track. LSMatch generated 147 total correspondences, out of which 129 were true positives, and 18 were false positives. In the Conference track, LSMatch achieved best precision. In recall, LSMatch is at par with other systems such as AMD and baseline matches.

2.3 Disease and Phenotype

This track is based on a real use case to find alignments between disease and phenotype ontologies. Specifically, the selected ontologies are the Human Phenotype Ontology (HPO), the Mammalian Phenotype Ontology (MP), the Human Disease Ontology (DOID), and the Orphanet and Rare Diseases Ontology (ORDO).

¹ <http://oaei.ontologymatching.org/2021/results/anatomy/index.html>

Table 1. Results of LSMatch on Anatomy track 2021

Matcher	Precision	Recall	F1	Runtime (sec)
LSMatch	0.997	0.618	0.763	98
ALIN	0.983	0.726	0.835	2190
ATMatcher	0.978	0.669	0.794	146
LogMapLite	0.962	0.728	0.828	2
AMD	0.96	0.739	0.835	3
Wiktionary	0.956	0.753	0.843	493
AML	0.956	0.927	0.941	32
Fine-TOM	0.933	0.808	0.866	15068
LogMap	0.917	0.848	0.881	7
TOM	0.916	0.794	0.851	2647
Lily	0.901	0.902	0.901	430
LogMapBio	0.874	0.914	0.894	1043
ALOD2Vec	0.828	0.766	0.796	261
OTMapOnto	0.646	0.811	0.72	16
Baseline	0.997	0.622	0.766	-

Table 2. Results of LSMatch on Conference track 2021

Matcher	Precision	Recall	F1
LSMatch	0.83	0.41	0.55
ATMatcher	0.69	0.51	0.59
LogMapLite	0.68	0.47	0.56
AMD	0.81	0.41	0.54
Wiktionary	0.66	0.53	0.59
AML	0.78	0.62	0.69
Fine-TOM	0.64	0.53	0.58
LogMap	0.76	0.56	0.64
TOM	0.69	0.48	0.57
Lily	0.62	0.43	0.51
ALOD2Vec	0.64	0.49	0.56
OTMapOnto	0.22	0.7	0.33
Baseline	0.76	0.41	0.53

Table 3 shows the performance of LSMatch system on disease and phenotype track at our end (official results are not available on the OAEI website at the time of submission of this paper). In the case of DOID-ORDO, LSMatch generated 1193 total correspondences, out of which 1178 were true positives and 15 were false positives. In the case of HP-MP, we received 685 total correspondences, out of which 683 were true positives, and 2 were false positives.

Table 3. The results of Disease and Phenotype track.

Test Case	Precision	Recall	F1	Runtime (sec)
doid-ordo	0.987426655	0.952303961	0.969547325	2004
hp-mp	0.997080292	0.981321839	0.989138306	2483

2.4 Common Knowledge Graphs

This track evaluates the ability of matching systems to map the schema (classes) of large common knowledge graphs such as DBpedia, YAGO and NELL. Publicly available knowledge graphs are highly complementary, and known for sharing data about real-world entities such as person, organization, and place. The goal of this task is to align classes from highly influential and domain-independent knowledge graphs. Table 4 shows the performance of LSMatch on Common Knowledge Graphs track. The evaluation was executed on a Linux virtual machine with 128 GB of RAM and 16 vCPUs (2.4 GHz) processors². LSMatch generated 102 total correspondences, out of which 101 were true positives, and 1 was false positive.

Table 4. The results of Common Knowledge Graphs track 2021

Matcher	Precision	Recall	F1	Runtime (sec)
LSMatch	0.99	0.78	0.87	1005
AML	0.0	0.0	0.0	319
LogMap	0.99	0.80	0.88	199
ALOD2Vec	1.0	0.80	0.89	253
OTMapOnto	0.90	0.84	0.87	496
KGMatcher	0.97	0.91	0.94	6935
Wiktionary	1.0	0.80	0.89	272
AMD	0.0	0.0	0.0	1107
ATmatcher	1.0	0.80	0.89	196
Baseline	1.0	0.60	0.75	37

² <http://oaei.ontologymatching.org/2021/results/commonKG/index.html>

2.5 Knowledge Graph

The Knowledge Graph Track contains isolated knowledge graphs with instance and schema data. The goal of the task is to match both the instances and the schema. The knowledge graphs were created in the course of the DBkWik by running the DBpedia extraction framework on Wikis from the Fandom Wiki hosting platform. Table 5 shows the performance of LSMatch on Knowledge Graph track. The evaluation was executed on a virtual machine (VM) with 32GB of RAM and 16 vCPUs (2.4 GHz) with Debian 9 operating system³. As LSMatch was only targeted for class matching. So, in Knowledge Graph track, LSMatch returned correspondences for classes only.

Table 5. The results of Knowledge Graph track 2021 (Classes only)

Matcher	Precision	Recall	F1
LSMatch	1.0	0.64	0.78
ALOD2Vec	1.0	0.67	0.80
AMD	0.40	0.18	0.25
AML	0.98	0.81	0.89
ATMatcher	0.97	0.79	0.87
BaselineAltLabel	1.0	0.59	0.74
BaselineLabel	1.0	0.59	0.74
Fine-TOM	1.0	0.66	0.80
KGMatcher	1.0	0.66	0.79
LogMap	0.93	0.71	0.81
OTMapOnto	0.59	0.64	0.61
TOM	1.0	0.71	0.83
Wiktionary	1.0	0.67	0.80

3 General Comments

The results show that the LSMatch system is performing at par with many of the systems that were tested at OAEI and is also performing better than some of them in some cases. As far as results are considered, we can see that the system achieves good precision in all the tracks. The current version lacks in recall, which affects the F1 score. The future iterations of the system will be targeted towards improving upon these measures.

The system has a lot of potentials to improve on different aspects such as multiple matchers can be employed together for finding missed out and tricky alignments; in the future versions, along with thesaurus.com, we can use knowledge bases like DBpedia, YAGO, or Wikidata as background knowledge, which can give more insights into the concepts represented into the ontologies for better alignments.

³ <http://oaei.ontologymatching.org/2021/results/knowledgegraph/index.html>

4 Conclusions

The LSMatch system is one of the good performers on multiple tracks. This year, the system was tested on 5 tracks, i.e., Anatomy, Conference, Disease and Phenotype, 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 string similarity measures.

References

1. Zhang, S., Hu, Y., and Bian, G. (2017, March). Research on string similarity algorithm based on Levenshtein Distance. In 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) (pp. 2247-2251). IEEE.
2. Hertling, S., Portisch, J., and Paulheim, H. (2019, September). Melt-matching evaluation toolkit. In International conference on semantic systems (pp. 231-245). Springer, Cham.
3. Shvaiko, P., and Euzenat, J. (2011). Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1), 158-176.
4. Nguyen, T. T. A., and Conrad, S. (2015, November). Ontology matching using multiple similarity measures. In 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K) (Vol. 1, pp. 603-611). IEEE.
5. <https://xlinux.nist.gov/dads/HTML/Levenshtein.html>
6. Aleksovski, Z., Ten Kate, W., and Van Harmelen, F. (2006, November). Exploiting the Structure of Background Knowledge Used in Ontology Matching. In *Ontology Matching* (p. 13).
7. thesaurus.com, URL: <https://www.thesaurus.com/>
8. Ochieng, P., and Kyanda, S. (2018). Large-scale ontology matching: state-of-the-art analysis. *ACM Computing Surveys (CSUR)*, 51(4), 1-35.
9. Anam, S., Kim, Y. S., Kang, B. H., and Liu, Q. (2015). Review of ontology matching approaches and challenges. *International journal of Computer Science and Network Solutions*, 3(3), 1-27.