LSMatch Results for OAEI 2021

Abstract
This paper presents a 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 four tracks, namely Anatomy, Conference, Disease and Phenotype, and Biodiversity and Ecology. LSMatch has demonstrated promising results in all four 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 the same by meaning but have different lexical representations [1]. This is our first OAEI participation, and we have targeted 4 tracks, i.e., Anatomy, Biodiversity and Ecology, Conference, and Disease and Phenotype. LSMatch system was wrapped using the MELT framework, and it is performing at par with some of the other systems in tracks and achieving best results in Disease and Phenotype track.

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 [2]. 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 schema (KS1 and KS2) as input to show the final set of alignments. LSMatch system takes input in any format and loaders convert input KS in the RDF graph.

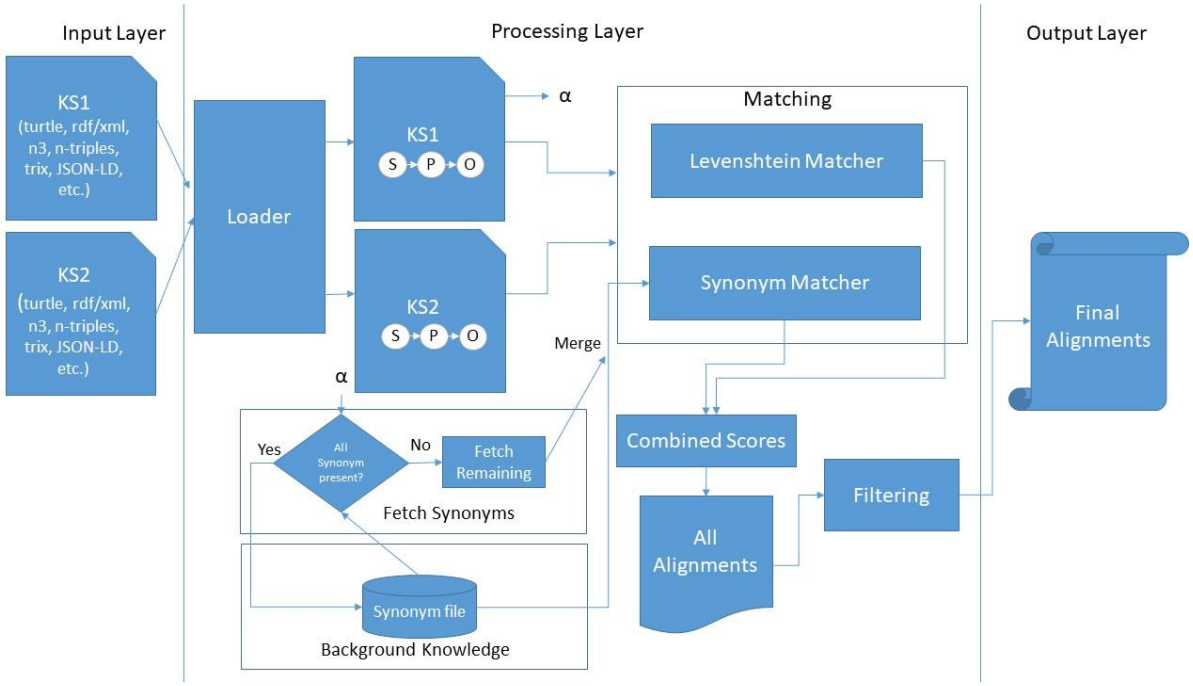

Fig. 1. Overview of LSMatch system

- Levenshtein matcher [3]: The LSMatch uses a string similarity matcher that calculates Levenshtein distance between the concepts. 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 Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other."
- Background knowledge [4]: To identify different lexical representations, LSMatch uses synonyms matcher that fetches synonyms from thesaurus.com by using their API [5]. For immediate availability of synonyms at the time of matching, we have pre-fetched the synonyms and kept them in json format.
- Synonym Matcher: We are fetching 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 is present or not. If some concept doesn't have synonyms pre-fetched for it, we are fetching them on the fly as well.

A dictionary stores information in <key, value> pairs where key is hashed [6-7]. LSMatch uses dictionary to store the alignments received from both the matchers along with the similarity score. As we target storing and updating the scores of pairs multiple times during the alignment process and having hashed key's allows us to do that efficiently. By default, we are keeping all the alignments with combined score (Levenshtein + Aynonym) of 0.5 for checking the alignments using variable thresholds. But the current version that we sent for testing has threshold 0.95 that gives us the final set of alignments.

2. Results

This section describes the results of the LSMatch system on four tracks namely: Anatomy, Conference, Disease and Phenotype, Biodiversity and Ecology. The evaluation was done on a server with 8GB RAM and CPU with 2.60GHz (6 Cores).

2.1 Anatomy

The Anatomy track consists of finding the alignments between the Adult Mouse Anatomy and the NCI Thesaurus describing the human anatomy. Table 1 shows the performance of LSMatch system on anatomy track. We received 940 total correspondences, out of which 937 were true positives and 3 were false positives.

Table 1. Results of LSMatch on Anatomy track 2021

Test Case	**Precision**	**Recall**	**F1**	**Runtime (sec)**
mouse-human-suite	0.996808511	0.618073879	0.763029316	84

2.2 Biodiversity and Ecology

This track consists on finding alignments between the Environment Ontology (ENVO) and the Semantic Web for Earth and Environment Technology Ontology (SWEET), and between the Flora Phenotype Ontology (FLOPO) and the Plant Trait Ontology (PTO). We tested our system only on FLOPO-PTO. Table 2 shows the performance of LSMatch system on biodiversity and ecology track. We received 140 total correspondences, out of which 140 were true positives and 0 were false positives.

Table 2. The results of Biodiversity and Ecology track 2021 on data sets FLOPO and PTO

Test Case	Precision	Recall	F1	Runtime (sec)
flopo-pto	1	0.573770492	0.729166667	348

2.3 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 3 (a) and (b) shows the performance of LSMatch system on conference track. We received 147 total correspondences, out of which 129 were true positives and 18 were false positives.

Table 3 (a) The results of Conference track 2021 based on combination of different ontologies

Test Case	**Precision**	**Recall**	**F1**	**Runtime**
ekaw-iasted	1	0.6	0.75	00:00:00
confof-edas	0.888889	0.421053	0.571429	00:00:00
confof-iasted	1	0.444444	0.615385	00:00:01
edas-ekaw	0.714286	0.217391	0.333333	00:00:01
cmt-ekaw	1	0.454545	0.625	00:00:01
edas-iasted	0.833333	0.263158	0.4	00:00:01
cmt-sigkdd	1	0.5	0.666667	00:00:00
iasted-sigkdd	0.916667	0.733333	0.814815	00:00:01
cmt-edas	1	0.615385	0.761905	00:00:01
confof-sigkdd	1	0.571429	0.727273	00:00:00
conference-iasted	0.8	0.285714	0.421053	00:00:01
cmt-conference	0.6	0.2	0.3	00:00:02
conference-ekaw	0.666667	0.32	0.432432	00:00:00
cmt-confof	0.8	0.25	0.380952	00:00:01
conference-edas	0.875	0.411765	0.56	00:00:01

conference-sigkdd	0.888889	0.533333	0.666667	00:00:00
conference-confof	0.875	0.466667	0.608696	00:00:01
cmt-iasted	0.8	1	0.888889	00:00:01
ekaw-sigkdd	1	0.636364	0.777778	00:00:00
confof-ekaw	0.888889	0.4	0.551724	00:00:00
edas-sigkdd	1	0.466667	0.636364	00:00:00

Table 3 (b) Aggregated results of LSMatch on Conference track 2021

Test Case	**Precision**	**Recall**	**F1**	**Runtime (sec)**
Aggregated	0.883219955	0.466249883	0.610315531	00:00:23

2.4 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). Table 4 shows the performance LSMatch system on disease and phenotype track. In the case of DOID-ORDO, we received 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 4. The results of Disease and Phenotype track 2021

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

3. General Comments

The results show that the LSMatch system is performing at par with many of the systems (either provide some reference for this statement or prove it here) that were tested at OAEI and is also performing better than some of them in some cases. As far as local results are considered, we can see that the system achieves good precision in all the tracks. The current version lacks recall, and we plan to improve in the future iterations of the system.

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 as well as instead of 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 ontology for better alignments.

4. Conclusions

The LSMatch system is one of the good performers on multiple tracks. This year, the system was tested on 4 tracks, i.e., Anatomy, Biodiversity and Ecology, Conference, and Disease and Phenotype. The system achieved considerably good precision in all the tracks but lacked behind in recall. We are planning to add 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. We are committed towards improving the system and we will keep addressing the issues it has.

References

[1] Zhang, S., Hu, Y., & 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] Shvaiko, P., & Euzenat, J. (2011). Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, *25*(1), 158-176.

[3] Xue, X., Wu, X., Jiang, C., Mao, G., & Zhu, H. (2021). Integrating sensor ontologies with global and local alignment extractions. *Wireless Communications and Mobile Computing*, *2021*.

[4] Aleksovski, Z., Ten Kate, W., & Van Harmelen, F. (2006, November). Exploiting the Structure of Background Knowledge Used in Ontology Matching. In *Ontology Matching* (p. 13).

[5] thesaurus.com, URL: https://www.thesaurus.com/

[6] Ochieng, P., & Kyanda, S. (2018). Large-scale ontology matching: state-of-the-art analysis. *ACM Computing Surveys (CSUR)*, *51*(4), 1-35.

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