ALOD2Vec Matcher Results for OAEI 2020

Jan Portisch[1,2][0000−0001−5420−0663], Michael Hladik[2][0000−0002−2204−3138], and Heiko Paulheim[1][0000−0003−4386−8195]

[1] Data and Web Science Group, University of Mannheim, Germany
{jan, heiko}@informatik.uni-mannheim.de
[2] SAP SE Product Engineering Financial Services, Walldorf, Germany
{jan.portisch, michael.hladik}@sap.com

Abstract. This paper presents the results of the *ALOD2Vec Matcher* in the *Ontology Alignment Evaluation Initiative* (OAEI) 2020. The matching system exploits a Web-scale dataset, i.e. *WebIsALOD*, as background knowledge source. In order to make use of the dataset, the *RDF2Vec* approach is applied to derive embeddings for each concept available in the dataset. *ALOD2Vec Matcher* participated in the OAEI 2018 campaign before. This is the system's second participation. The matching system has been extended, improved, and achieves better results this year.

Keywords: Ontology Matching · Ontology Alignment · External Resources · Background Knowledge · Knowledge Graph Embeddings · RDF2Vec

1 Presentation of the System

1.1 State, Purpose, General Statement

The *ALOD2Vec Matcher* is an element-level, label-based matcher which uses a large-scale Web-crawled RDF dataset of hypernymy relations as general purpose background knowledge. The dataset contains many tail-entities as well as instance data such as persons or places which cannot be found in common thesauri. In order to exploit the external dataset, a neural language model approach is used to obtain a vector for each concept contained in the dataset. This matching system system was initially introduced at the OAEI 2018 [14] and has been completely re-implemented. The implementation is now based on the *Matching EvaLuation Toolkit* [5,11] as well as the KGvec2go [12] REST API. A contribution of this paper is also an extension to the MELT framework in the form of a KGvec2go Java client available in the MELT-ML module [6] of MELT 2.6.

1.2 Specific Techniques Used

After the basic concepts of this matcher are introduced (*Foundations*), the specific techniques applied are presented.

Foundations

WebIsALOD Dataset A frequent problem that occurs when working with external background knowledge is the fact that less common entities are not contained within a knowledge base. The *WebIsA* [17] database is an attempt to tackle this problem by providing a dataset which is not based on a single source of knowledge – like *DBpedia* [8] – but instead on the whole Web: The dataset consists of hypernymy relations extracted from the *Common Crawl*[3], a freely downloadable crawl of a significant portion of the Web. A sample triple from the dataset is *european_union skos:broader international_organization*[4]. The dataset is also available via a Linked Open Data (LOD) endpoint[5] under the name *WebIsALOD* [4]. In the LOD dataset, a machine-learned confidence score $c \in [0, 1]$ is assigned to every hypernymy triple indicating the assumed degree of truth of the statement.

RDF2Vec The background dataset can be viewed as a very large knowledge graph; in order to obtain a similarity score for nodes and edges in that graph, the *RDF2Vec* [16] approach is used. It applies the *word2vec* [9,10] model to RDF data: Random walks are performed for each node and are interpreted as sentences. After the walk generation, the sentences are used as input for the word2vec algorithm. As a result, one obtains a vector for each word, i.e., a concept in the RDF graph. Multiple flavors of RDF2Vec have been developed in the past such as biased walks [1] or *RDF2Vec Light* [13].[6]

KGvec2go Training embeddings on large knowledge graphs can be computationally very expensive. Moreover, the resulting embedding models can be very large since a multidimensional vector needs to be persisted for every node in the knowledge graph. However, most downstream applications require only a small subset of node vectors. The *KGvec2go* project [12] addresses these problems by providing a free REST API[7] for pre-trained RDF2Vec models on various large knowledge graphs (among which WebIsALOD is also available).

Monolingual Matching *ALOD2Vec Matcher* is a monolingual matching system. For the alignment process, the system retrieves the labels of all elements of the ontologies to be matched. A filter adds all simple string matches to the final alignment in order to increase the performance. The remaining labels are linked to concepts in the background dataset, are compared, and the best solution is added to the final alignment. A high-level view of the matching system is provided in Figure 1.

[3] see `http://commoncrawl.org/`
[4] see `http://webisa.webdatacommons.org/concept/european_union_`
[5] see `http://webisa.webdatacommons.org/`
[6] For a good overview of the RDF2Vec approach and its applications, refer to `http://www.rdf2vec.org/`
[7] see `http://kgvec2go.org/api.html`

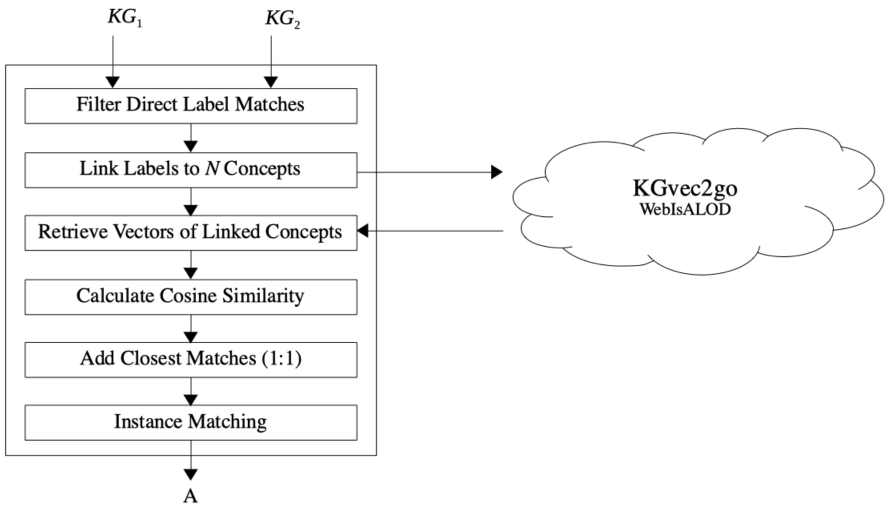

Fig. 1. High-level view of the ALOD2Vec matching process. KG_1 and KG_2 represent the input ontologies and optionally instances. The final alignment is referred to as A.

The first step is to link the obtained labels from the ontology to concepts in the WebIsALOD dataset. Therefore, string operations are performed on the label and it is checked whether the label is available in WebIsALOD. If it cannot be found, a token-lookup is performed. Given two entities e_1 and e_2, the matcher uses their textual labels to link them to concepts e'_1 and e'_2 in the external dataset. Afterwards, the embedding vectors $v_{e'_1}$ and $v_{e'_2}$ of the linked concepts (e'_1 and e'_2) are retrieved via a Web request and the cosine similarity between those is calculated. Hence: $sim(e_1, e_2) = sim_{cosine}(v_{e'_1}, v_{e'_2})$. If $sim(e_1, e_2) > t$ where t is a threshold in the range of 0 and 1, a correspondence is added to a temporary alignment. In a last step, a one-to-one arity is enforced by applying a *Maximum Weight Bipartite* [2] filter on the temporary alignment.

In order to consume the vectors in Java, a client has been implemented and contributed to the MELT-ML module. The KGvec2go REST API can now be accessed though class `KGvec2goClient`. Even though this matcher only uses the WebIsALOD dataset, the implementation supports all datasets accessible on KGvec2go. The extension is available by default in MELT 2.6.

Instance Matching For the 2020 version of the matching system, an instance matching module has been added. After classes and properties have been matched, instances are matched using a string index. The confidence score assigned to instances belonging to matched classes is higher than that of matches between instances belonging to non-matched classes.

Explainability *ALOD2Vec Matcher* provices an explanation for every correspondence that is added to the final alignment. Therefore, the extension capabilities of the alignment format [3] are used. Two concrete examples from the *Anatomy track* for explanations of the matching system are: "Label 'aqueous humour' of ontology 1 and label 'Aqueous Humor' of ontology 2 have a very similar writing." or "The following two label sets have a cosine above the given threshold: |lens|anterior|epithelium| and |anterior|surface|lens|". In order to explain a correspondence, the `description` property[8] of the *Dublin Core Metadata Initiative* is used.

1.3 Extensions to the Matching System for the 2020 Campaign

The 2020 system has been completely rewritten. Among the significant changes are an improved handling of string matches, an instance matching module for the *knowledge graph track* [7], explanations on the level of correspondences, a simplified linking process as well as the usage of a Web endpoint compared to a local key value database that has been used before. It is important to note that the 2020 system uses the KGvec2go model for ALOD2Vec which is not equal to the model trained in 2018. Due to the usage of the KGvec2go API, the SEALS package is now several magnitudes smaller than before in terms of required disk space.[9] The smaller package cost comes at the price of a slower system runtime due to API calls. However, this matcher still scored at the exact median of all matching systems in terms of runtime on the anatomy track this year. The 2020 implementation is publicly available on GitHub.[10]

2 Results

2.1 Anatomy Track

On the anatomy dataset, the recall could be significantly improved in 2020 compared to the 2018 version of the matching system. Despite a drop in precision, the new *ALOD2Vec Matcher* achieves an overall higher F_1 score. Due to multiple API calls to KGvec2go, the runtime performance decreased compared to the 2018 version of the matcher.

2.2 Conference Track

On the conference track, the new matcher configuration achieved a better result than the 2018 one in terms of F_1.

[8] see `http://purl.org/dc/terms/description`

[9] The 2018 version of the matching system had to be submitted via a download link due to its large size. The 2020 version was submitted using the default process.

[10] see `https://github.com/janothan/ALOD2VecMatcher`

2.3 Knowledge Graph Track

This is the first year that *ALOD2Vec Matcher* participates in the knowledge graph track. Due to the new instance matching module, this matcher obtains the second best results achieving almost the same score as the *Wiktionary Matcher 2020* [15].

3 Conclusion

In this paper, we presented the newest version of the *ALOD2Vec Matcher*, a matcher utilizing an RDF2Vec vector representation of the *WebIsALOD* dataset, as well as its results in the 2020 OAEI. Overall, the results of the matching system could be significantly improved compared to its last OAEI participation.

References

1. Cochez, M., Ristoski, P., Ponzetto, S.P., Paulheim, H.: Biased graph walks for RDF graph embeddings. In: Akerkar, R., Cuzzocrea, A., Cao, J., Hacid, M. (eds.) Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, WIMS 2017, Amantea, Italy, June 19-22, 2017. pp. 21:1–21:12. ACM (2017). https://doi.org/10.1145/3102254.3102279, https://doi.org/10.1145/3102254.3102279
2. Cruz, I.F., Antonelli, F.P., Stroe, C.: Efficient selection of mappings and automatic quality-driven combination of matching methods. In: Proceedings of the 4th International Conference on Ontology Matching-Volume 551. pp. 49–60. Citeseer (2009)
3. David, J., Euzenat, J., Scharffe, F., dos Santos, C.T.: The alignment API 4.0. Semantic Web **2**(1), 3–10 (2011). https://doi.org/10.3233/SW-2011-0028, https://doi.org/10.3233/SW-2011-0028
4. Hertling, S., Paulheim, H.: Webisalod: Providing hypernymy relations extracted from the web as linked open data. In: d'Amato, C., Fernández, M., Tamma, V.A.M., Lécué, F., Cudré-Mauroux, P., Sequeda, J.F., Lange, C., Heflin, J. (eds.) The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II. Lecture Notes in Computer Science, vol. 10588, pp. 111–119. Springer (2017). https://doi.org/10.1007/978-3-319-68204-4_11, https://doi.org/10.1007/978-3-319-68204-4_11
5. Hertling, S., Portisch, J., Paulheim, H.: MELT - Matching EvaLuation Toolkit. In: Semantics 2019 SEM2019 Proceedings. Karlsruhe (2019, to appear)
6. Hertling, S., Portisch, J., Paulheim, H.: Supervised ontology and instance matching with MELT. In: OM@ISWC 2020 (2020), to appear
7. Hofmann, A., Perchani, S., Portisch, J., Hertling, S., Paulheim, H.: Dbkwik: Towards knowledge graph creation from thousands of wikis. In: Nikitina, N., Song, D., Fokoue, A., Haase, P. (eds.) Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 23rd - to - 25th, 2017. CEUR Workshop Proceedings, vol. 1963. CEUR-WS.org (2017), http://ceur-ws.org/Vol-1963/paper540.pdf

8. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**(2), 167–195 (2015). https://doi.org/10.3233/SW-140134, `https://doi.org/10.3233/SW-140134`
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (eds.) 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings (2013), `http://arxiv.org/abs/1301.3781`
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States. pp. 3111–3119 (2013), `http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality`
11. Portisch, J., Hertling, S., Paulheim, H.: Visual analysis of ontology matching results with the melt dashboard. In: The Semantic Web: ESWC 2020 Satellite Events (2020)
12. Portisch, J., Hladik, M., Paulheim, H.: Kgvec2go - knowledge graph embeddings as a service. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020. pp. 5641–5647. European Language Resources Association (2020), `https://www.aclweb.org/anthology/2020.lrec-1.692/`
13. Portisch, J., Hladik, M., Paulheim, H.: Rdf2vec light - a lightweight approach for knowledge graph embeddings. In: Proceedings of the ISWC 2020 Posters Demonstrations (2020), to appear
14. Portisch, J., Paulheim, H.: Alod2vec matcher. In: OM@ISWC. CEUR Workshop Proceedings, vol. 2288, pp. 132–137. CEUR-WS.org (2018)
15. Portisch, J., Paulheim, H.: Wiktionary Matcher results for OAEI 2020. In: OM@ISWC 2020 (2020), to appear
16. Ristoski, P., Rosati, J., Noia, T.D., Leone, R.D., Paulheim, H.: Rdf2vec: RDF graph embeddings and their applications. Semantic Web **10**(4), 721–752 (2019). https://doi.org/10.3233/SW-180317, `https://doi.org/10.3233/SW-180317`
17. Seitner, J., Bizer, C., Eckert, K., Faralli, S., Meusel, R., Paulheim, H., Ponzetto, S.P.: A large database of hypernymy relations extracted from the web. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016. European Language Resources Association (ELRA) (2016), `http://www.lrec-conf.org/proceedings/lrec2016/summaries/204.html`