ALOD2vec Matcher Results for OAEI 2021

Jan Portisch[1,2][0000−0001−5420−0663] 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 Business Technology Platform - One Domain Model, Walldorf, Germany
jan.portisch@sap.com

Abstract. This paper presents the results of the *ALOD2vec Matcher* in the *Ontology Alignment Evaluation Initiative* (OAEI) 2021. 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 and 2020 campaigns before. This is the system's third participation.[3]

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 [13] and also participated in the 2020 campaign [10]. The implementation is based on the *Matching EvaLuation Toolkit* [6] as well as the *KGvec2go* [11] REST API to obtain vector representations via a Web API.

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* [16] database is an attempt to tackle this problem by providing a dataset which is not based on a single source of knowledge – like *DBpedia* [7] – but instead on the whole Web: The dataset consists of hypernymy relations extracted from the *Common Crawl*[4], a freely downloadable crawl of a significant portion of the Web. A sample triple from the dataset is *european_union skos:broader international_organization*[5]. The dataset is also available via a Linked Open Data (LOD) endpoint[6] under the name *WebIsALOD* [5]. 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* [15] approach is used. It applies the *word2vec* [8,9] 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* [12].[7]

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 [11] addresses these problems by providing a free REST API[8] 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.

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

[4] see `http://commoncrawl.org/`
[5] see `http://webisa.webdatacommons.org/concept/european_union_`
[6] see `http://webisa.webdatacommons.org/`
[7] For a good overview of the RDF2vec approach and its applications, refer to `http://www.rdf2vec.org/`
[8] see `http://kgvec2go.org/api.html`

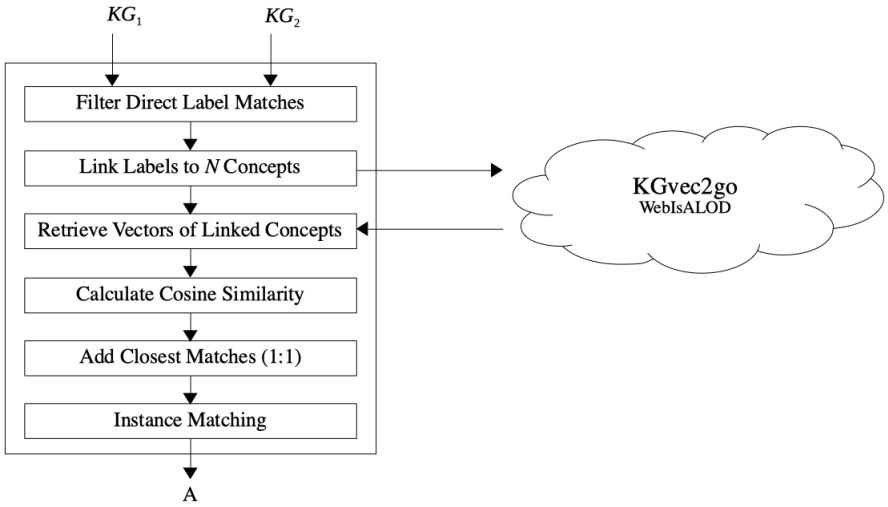

Fig. 1. High-level view of the ALOD2vec matching process. KG_1 and KG_2 represent the input ontologies and optionally instances. The final alignment is referred to as A.

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 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* provides 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[9] of the *Dublin Core Metadata Initiative* is used.

1.3 Extensions to the Matching System for the 2021 Campaign

For the 2021 campaign, the matching system was adapted to use the latest MELT release and was packaged as MELT Web Docker[10] container. The 2021 implementation is publicly available on GitHub.[11]

2 Results

2.1 Anatomy Track

On the anatomy dataset, the system scores a precision of 0.828, a recall of 0.766, and an F_1 of 0.796.

2.2 Conference Track

On the conference track, the matcher achieves a recall of 0.49 and a precision of 0.64. The overall F_1 score on `ra1-M3` was 0.59.

2.3 Multifarm Track

Since the *WebIsALOD* dataset is only available in English, the focus of the *ALOD2vec Matcher* is on monolingual matching tasks.

2.4 LargeBio Track

In its current version, the LargeBio track is too large for the matching system's architecture. There is a tradeoff in package size and runtime performance (a large package with all vectors matches faster than the submitted small package which obtains vectors at runtime from KGvec2go). The current architecture of *ALOD2vec Matcher* is not intended for large-scale matching – however, the matching algorithm itself could be used for large-scale matching.

2.5 Knowledge Graph Track

The system could complete all matching tasks in time. As in the previous year, this matcher obtains the second best results achieving almost the same score as the *Wiktionary Matcher 2021* [14]. The overall F_1 score was 0.87 on the complete track.

[9] see `http://purl.org/dc/terms/description`
[10] see `https://dwslab.github.io/melt/matcher-packaging/web`
[11] see `https://github.com/janothan/ALOD2VecMatcher`

2.6 Common Knowledge Graph Track

This year, a new track was added to the OAEI: The *Common Knowledge Graph Track* [4]. Although not optimized for this track, *Alod2vec Matcher* achieved the second best result with an F_1 score of 0.89.

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 2021 OAEI. In the future, the matching system could be improved by using another, potentially larger or newer, hypernymy database, by exploiting other embedding algorithms, and by adding further matching strategies to the overall algorithms such as checking of logical constraints.

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
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
4. Fallatah, O., Zhang, Z., Hopfgartner, F.: A gold standard dataset for large knowledge graphs matching. In: Shvaiko, P., Euzenat, J., Jiménez-Ruiz, E., Hassanzadeh, O., Trojahn, C. (eds.) Proceedings of the 15th International Workshop on Ontology Matching co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 2, 2020. CEUR Workshop Proceedings, vol. 2788, pp. 24–35. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2788/om2020_LTpaper3.pdf
5. 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
6. Hertling, S., Portisch, J., Paulheim, H.: MELT - matching evaluation toolkit. In: Acosta, M., Cudré-Mauroux, P., Maleshkova, M., Pellegrini, T., Sack, H., Sure-Vetter, Y. (eds.) Semantic Systems. The Power of AI and Knowledge Graphs - 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9-12, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11702, pp. 231–245. Springer (2019), https://doi.org/10.1007/978-3-030-33220-4_17

7. 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
8. 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
9. 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)
10. Portisch, J., Hladik, M., Paulheim, H.: Alod2vec matcher results for OAEI 2020. In: Shvaiko, P., Euzenat, J., Jiménez-Ruiz, E., Hassanzadeh, O., Trojahn, C. (eds.) Proceedings of the 15th International Workshop on Ontology Matching co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 2, 2020. CEUR Workshop Proceedings, vol. 2788, pp. 147–153. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2788/oaei20_paper2.pdf
11. 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/
12. Portisch, J., Hladik, M., Paulheim, H.: Rdf2vec light - A lightweight approachfor knowledge graph embeddings. In: Taylor, K.L., Gonçalves, R.S., Lécué, F., Yan, J. (eds.) Proceedings of the ISWC 2020 Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 19th International Semantic Web Conference (ISWC 2020), Globally online, November 1-6, 2020 (UTC). CEUR Workshop Proceedings, vol. 2721, pp. 79–84. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2721/paper520.pdf
13. Portisch, J., Paulheim, H.: Alod2vec matcher. In: Shvaiko, P., Euzenat, J., Jiménez-Ruiz, E., Cheatham, M., Hassanzadeh, O. (eds.) Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018. CEUR Workshop Proceedings, vol. 2288, pp. 132–137. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2288/oaei18_paper3.pdf
14. Portisch, J., Paulheim, H.: Wiktionary Matcher results for OAEI 2021. In: OM@ISWC 2021 (2021), to appear
15. 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
16. 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