ALOD2Vec Matcher*

Jan Portisch[0000−0001−5420−0663] and Heiko Paulheim[0000−0003−4386−8195]

Data and Web Science Group, University of Mannheim, Germany
jan.portisch@sap.com, heiko@informatik.uni-mannheim.de

Abstract. In this paper, we introduce the *ALOD2Vec Matcher*, an ontology matching tool that exploits a Web-scale data set, i.e., *WebIsA-LOD*, as external knowledge source. In order to make use of the data set, the *RDF2Vec* approach is chosen to derive embeddings for each concept available in the data set.
We show that it is possible to use very large RDF graphs as external background knowledge source for the task of ontology matching.

Keywords: Ontology Matching · Ontology Alignment · External Resources · Vector Space 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 data set of hypernymy relations as background knowledge. One advantage of that data set is the inclusion of many tail-entities, as well as instance data, such as persons or places, which cannot be found in thesauri. In order to make use of the external data set, a neural language model approach is used to calculate an embedding vector for each concept contained in it.
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 data set. Then, the pre-calculated embedding vectors $v_{e_1'}$ and $v_{e_2'}$ of the linked concepts (e_1' and e_2') are retrieved and the cosine similarity between those is calculated. Hence: $sim(e_1, e_2) = sim_{cosine}(v_{e_1'}, v_{e_2'})$.
The resulting alignment is homogenous, i.e., classes, object properties, and datatype properties are handled separately. In addition, the matcher enforces a one-to-many matching restriction.

1.2 Specific techniques used

For the alignment process, the matcher retrieves textual descriptions 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 of the background data set, are compared, and the best solution is added to the final alignment. A high-level view of the system is depicted in figure 1.

* Supported by SAP SE.

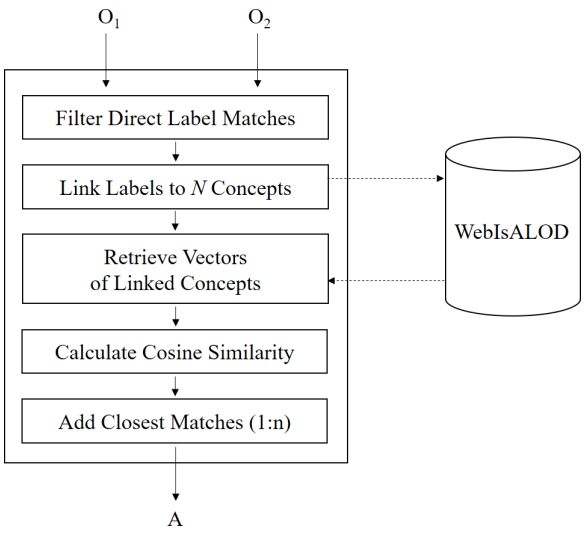

Fig. 1. ALOD2Vec Matching Process

WebIsALOD Data Set When working with knowledge bases in order to exploit the contained knowledge in applications, a frequent problem is the fact that less common entities are not contained within the knowledge base. The *WebIsA* [7] database is an attempt to tackle this problem by providing a data set which is not based on a single source of knowledge – like DBpedia [3] – but instead on the whole Web: The data set consists of hypernymy relations extracted from the *Common Crawl*[1], a freely downloadable crawl of a significant portion of the Web. A sample triple from the data set is *european_union skos:broader international_organization*[2]. The data set is also available via a Linked Open Data (LOD) endpoint[3] under the name *WebIsALOD* [2]. In the LOD data set, 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 data set can be viewed as a very large knowledge graph; in order to obtain a similarity score for nodes in that graph, the *RDF2Vec* [6] approach is used. It applies the *word2vec* [4,5] 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. The approach is used here to obtain vectors for all concepts in the WebIsALOD data set.

[1] see `http://commoncrawl.org/`
[2] see `http://webisa.webdatacommons.org/concept/european_union_`
[3] see `http://webisa.webdatacommons.org/`

Linking The first step is to link the obtained labels from the ontology to concepts in the WebIsALOD data set. Therefore, string operations are performed on the label and it is checked whether the label is available in WebIsALOD. If it cannot be found, labels consisting of multiple words are truncated from the right, and the process is repeated to check for sub-concepts. For example, the label *United Nations Peacekeeping Mission in Mali* cannot be found in WebIsALOD. Therefore, it is truncated until the longest label from the left is found – in this case *United Nations*. The process is repeated until all tokens are processed. The resulting concepts for the given label are: *United Nations*[4], *peacekeeping mission*[5], and *Mali*[6].

Similarity Calculation As stated before, labels are linked to concepts, their vectors are retrieved, and the cosine similarity between them is used as similarity score.

There are cases in which parts of a label cannot be found, however, for example in *tubule macula* and in *macula lutea* both times only *macula* can be found using the WebIsALOD data set. If only the found concepts would be used to calculate the similarity between the concepts, a perfect score would be obtained because $sim(macula, macula) = 1.0$. This is not precise as the approach does not allow to discriminate between perfect matches due to incomplete linking and *real* perfect matches. Therefore, a penalty factor $p \in [0, 1]$ is introduced that is to be multiplied with the final similarity score and which lowers the score for incomplete links; $p = 0$ indicates the maximal penalty, $p = 1$ indicates no penalty. The calculation of p is depicted in equation 1:

$$p = 0.5 * \frac{|Found\ Concepts\ L_1|}{|Possible\ Concepts\ L_1|} + 0.5 * \frac{|Found\ Concepts\ L_2|}{|Possible\ Concepts\ L_2|} \quad (1)$$

where L_1 is the label of the first concept and L_2 is the label of the second one; $|Found\ Concepts\ L_i|$ is the number of tokens for which a concept could be found (minus stopwords) and $|Possible\ Concepts\ L_i|$ is the number of tokens of the label without stopwords. The penalty score is multiplied with the final similarity score. Hence, incomplete linkages are penalized.

If two labels were matched to multiple concepts, a resolution is required. In this case the best average similarity is used:

$$sim_{average} = \frac{\Sigma_{i \in c_1}^{|c_1|} Max_{j \in c_2}^{|c_2|} sim(c_{1_i}, c_{2_j})}{|c_1|} \quad (2)$$

where c_1 and c_2 represent two individual concepts and c_{1_i}, respectively c_{2_j}, represent the i^{th} and j^{th} sub-concept of c_1 and c_2; $|c_1|$ and $|c_2|$ are the number of sub-concepts of c_1 and c_2; c_1 is the concept with more tokens.

[4] see `http://webisa.webdatacommons.org/concept/united_nations_`
[5] see `http://webisa.webdatacommons.org/concept/peacekeeping_mission_`
[6] see `http://webisa.webdatacommons.org/concept/_mali_`

Typically, there is more than one label to an entity of an ontology. Therefore, a score-matrix is used: Every label of an entity is linked and compared to every label of the other entity and the best score is returned.

RDF2Vec Configuration Parameters We generated 100 sentences of depth 8 for each node in the WebIsALOD data set for the training process of the model. In order to have also sentences for nodes that do not have out-going edges, those were identified and sentences were generated backwards and afterwards reversed. The sentences were generated in a biased fashion [1], i.e., high-confidence edges are followed with a higher probability. Eventually, the embeddings were trained using the continuous bag of words (CBOW) approach with the parameters of the original RDF2Vec paper: $window\ size = 5$, $number\ of\ iterations = 5$, $negative\ sampling = true$, $negative\ samples = 25$, $average\ input\ vector = true$, and 200 dimensional embeddings.

2 Results

2.1 Anatomy

For the Anatomy data set, the matcher achieves a higher recall and F_1 score compared to the baseline solution. However, the true positives are mostly exact lexical matches or share many common tokens.
Concerning runtime-performance, *ALOD2Vec Matcher* performs in the upper half of all matchers that participated in the Anatomy track.

2.2 Conference

On the Conference data set, it can be seen that the matcher is better in aligning classes than in aligning properties. This is in line with the results reported for other matchers. In this case, it is due to fewer lexical matches in properties as well as the higher usage of non-nouns which cannot be properly linked to the background knowledge source.

2.3 Large BioMed

For the Large BioMed matching tasks, the matcher is capable of aligning the small fragments within the given time frame of 6 hours. While *ALOD2Vec Matcher* performs slightly above the 2017 and 2018 F_1 averages on the small FMA-NCI data set, it perfoms in the lower half for the remaining ones.

2.4 Complex Track

Although the matcher presented here is not capable of generating complex correspondences yet, it could produce results for the entity identification subtask for two data sets: On GeoLink, *ALOD2Vec Matcher* achieved the highest F_1 score and recall of all matchers that participated; on Hydrograph, alignments for the English ontologies could be generated and scored within the median.

3 General Comments

3.1 Comments on the results

The matcher performs above the given baselines. However, the matches are still rather trivial and mostly share common tokens.

There are multiple reasons for the mediocre performance. First, the underlying data set is very noisy: It contains a lot of wrong information (e.g. *fish skos:broader fisher*)[7], subjective information (e.g. *donald_trump skos:broader lunatic*)[8], and is not strictly hierarchical (e.g. *live skos:broader quality*, and vice versa)[9]. In addition, the tail-entity problem is still not solved because very specific entities are involved in very few hypernymy statements and their resulting vectors are likely not meaningful (e.g. *complex congenital heart defect*)[10].

Besides the pitfalls of the data set, the matcher cannot handle homonyms, non-nouns, or non-English labels.

3.2 Discussions on the way to improve the proposed system

There are three ways in which the current research focusing on this approach can be improved in the future: Firstly, more propositionalization techniques for very large data sets could be explored. Secondly, the matcher itself can be enhanced to use more information available in ontologies such as their structure. And lastly, the data sets to be used can be improved. WebIsALOD is only one Web-scale RDF data set and still has some pitfalls such as the restriction to hypernymy relations and noise. More such data sets can be created and used in the future.

4 Conclusion

In this paper, we presented the *ALOD2Vec Matcher*, a matcher utilizing a Web-crawled knowledge data set by applying the RDF2Vec methodology to a hypernymy data set extracted from the Web. It could be shown that it is possible to use very large RDF graphs as external background knowledge and the RDF2Vec methodology for the task of ontology matching.

References

1. Cochez, M., Ristoski, P., Ponzetto, S.P., Paulheim, H.: Biased graph walks for rdf graph embeddings. In: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics. p. 21. ACM (2017)

[7] see `http://webisa.webdatacommons.org/concept/_fish_`
[8] see `http://webisa.webdatacommons.org/concept/donald_trump_`
[9] see `http://webisa.webdatacommons.org/concept/_quality_`
[10] see `http://webisa.webdatacommons.org/concept/complex+congenital+heart_defect_`

2. Hertling, S., Paulheim, H.: WebIsALOD: Providing Hypernymy Relations Extracted From the Web as Linked Open Data. In: International Semantic Web Conference. pp. 111–119. Springer (2017)
3. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Bizer, C.: DBpedia A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web **6**(2), 167–195 (2012)
4. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781 (2013)
5. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed Representations of Words and Phrases and Their Compositionality. In: Advances in Neural Information Processing Systems. pp. 3111–3119 (2013), `http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality`
6. Ristoski, P., Rosati, J., Di Noia, T., De Leone, R., Paulheim, H.: RDF2vec: RDF Graph Embeddings and Their Applications. Semantic Web Journal (2017), `http://www.semantic-web-journal.net/system/files/swj1495.pdf`
7. 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: LREC (2016)