The Supervised Semantic Similarity Toolkit

Rita T. Sousa, Sara Silva, and Catia Pesquita

LASIGE, Faculdade de Ciências da Universidade de Lisboa
{risousa,sgsilva,clpesquita}@ciencias.ulisboa.pt

Abstract. Knowledge graph-based semantic similarity measures have been used in several applications. Although knowledge graphs typically describe entities according to different semantic aspects modeled in ontologies, state-of-the-art semantic similarity measures are general-purpose since they consider the whole graph or depend on expert knowledge for fine-tuning.
We present a novel toolkit that can tailor aspect-oriented semantic similarity measures to fit a particular view on similarity. It starts by identifying the semantic aspects, then computes similarities for each semantic aspect, and finally uses a supervised machine learning method to learn a supervised semantic similarity according to the similarity proxy. The toolkit combines six taxonomic semantic similarity and four embedding similarity measures and provides baseline evaluation approaches.
This extended abstract is related to the paper "Towards Supervised Biomedical Semantic Similarity" accepted to the SeWeBMeDA 2022 but focuses on our work's technical contribution whereas the workshop submission focuses on the use case for biomedical informatics.

Keywords: Ontology · Knowledge Graph · Graph Embedding · Semantic Similarity · Machine Learning.

1 Introduction

Semantic similarity between entities in knowledge graphs (KGs) is essential for several tasks, especially in data mining and machine learning. State-of-the-art semantic similarity measures (SSMs), both taxonomic and graph embedding-based measures, are general-purpose and either consider the graph as a whole or depend on domain expert knowledge. However, KGs provide multiple semantic aspects (SAs) (Definition 1) or perspectives over an entity and, depending on our viewpoint of the domain, different SAs should be considered in similarity computation. In previous work, we developed a methodology to predict protein interactions that uses genetic programming, a machine learning (ML) method, to evolve combinations of aspect-oriented semantic similarities [8]. The positive results inspired us to hypothesize that, not only in the biomedical domain, if data regarding a similarity proxy (Definition 2) is available, we can learn a supervised semantic similarity tailored to capture a specific similarity view that combines different SAs.

Definition 1. *A **semantic aspect** represents a perspective of the representation of KG entities. It can correspond to portions of the graph (e.g., describing a protein only through the biological process subgraph of the Gene Ontology) or a given set of property types (e.g., describing a person only through properties having geographical locations as a range).*

Definition 2. *A **similarity proxy** is an estimation of the similarity between two entities that relies on objective representations of entities and calculate similarity using mathematical expressions or other algorithms.*

We have developed a toolkit[1] that learns a supervised semantic similarity between entities represented in KGs tailored towards a specific similarity proxy. This tailoring is achieved by using supervised ML methods where the input values are the similarities for different SAs, and the expected outputs are the proxy similarity values. Currently, our toolkit supports 10 SSMs (4 based on KG embeddings and 6 based on taxonomic similarity) coupled with 8 ML methods (classical ML approaches and neural network-based approaches). Since our toolkit is especially suited to KGs with several SAs, such as biomedical KGs, we applied it in a collection of benchmark datasets for KG-based similarity in the biomedical domain [2]. It is, however, domain-independent and readily applicable to other applications, such as recommender systems where the similarity computation between users is essential.

2 The toolkit

Our toolkit, shown in Figure 1, needs a KG and a list of instance pairs with proxy similarity values and is able to: (1) identify the SAs that describe the KG entities (2) compute KG-based similarities according to different SAs and using different SSMs; (3) train supervised ML algorithms to learn a supervised semantic similarity according to the similarity proxy for which we want to tailor the similarity; (4) evaluate the supervised semantic similarity against a set of baselines. This framework is independent of the SAs, the specific implementation of KG-based similarity and the ML algorithm employed in supervised learning.

Semantic aspects selection In this work, we consider KGs where real-word entities are annotated with classes from ontologies. Ontologies structure their classes and the relationships between them as a directed acyclic graph. A semantic annotation is about assigning real-world entities to ontology classes describing them. Therefore, our toolkit takes as input an ontology file and an instance annotation file to generate the KG, where the nodes represent ontology classes and real-world entities, and edges are employed in representing ontology classes' relations and semantic annotations.

As default, our toolkit uses subgraphs rooted in the classes at a distance of one from the KG root class or the subgraphs when the KGs have multiple roots

[1] https://github.com/liseda-lab/Supervised-SS

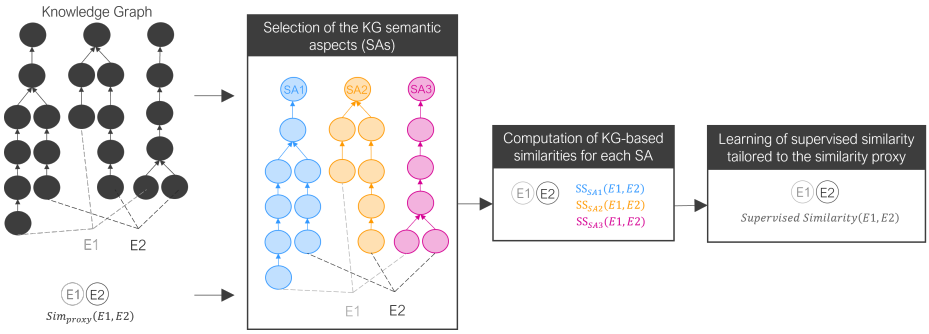

Fig. 1. Overview of the proposed framework. Considering a KG with its three SAs, each instance of the dataset representing a pair between an entity E1 and an entity E2 is characterized by three SS values corresponding to the semantic similarity between them for the three SAs, and a proxy similarity value. The models returned in the second step are then the combinations of the similarity scores of the three SAs.

as SAs. However, SAs can also be manually defined by selecting the root classes that anchor the aspects.

Similarity computation for each semantic aspect For the computation of KG-based similarities for each SA, our toolkit employs 10 KG-based SSMs.

Taxonomic semantic similarity The taxonomic semantic similarity can be calculated using six state-of-the-art measures, from combining two information content (IC) approaches (IC_{Seco}, $\text{IC}_{\text{Resnik}}$) with one of three set similarity measures (ResnikBMA, ResnikMax, SimGIC). IC is a measure of how specific and informative a concept is, giving SSMs the ability to weigh the similarity of two concepts according to their specificity. IC_{Resnik} [5] is an extrinsic IC based on the number of occurrences of a concept in a corpus of texts. IC_{Seco} [7] is a structure-based approach based on structural information extracted from the ontology, namely the number of direct and indirect descendants.
Two types of approaches can be employed to calculate semantic similarity for two entities, each annotated with a set of concepts: pairwise approaches, where pairwise comparisons between all concepts annotating each entity are considered, or groupwise approaches. Resnik [5] is a pairwise class-based measure in which the similarity between two classes corresponds to the IC of their most informative common ancestor. Pairwise scores are then summarised using an aggregation strategy. For *ResnikBMA*, only the best-matching pair for each term is considered. For *ResnikMax*, the maximum of the pairwise similarities is used instead. *SimGIC* [4] is a groupwise approach based on a Jaccard index in which each term is weighted by its IC.

Graph embedding similarity Four different representative graph embedding approaches can be employed to generate graph embeddings. *TransE* [1] is a trans-

lational distance approach, where each fact represents the distance between the two entities after a translation carried out by the relations. *distMult* [9] is a semantic matching approach that exploits similarity-based scoring functions by matching latent semantics of entities and relations embodied in their vector space representations. *RDF2Vec* [6] is a path-based approach that performs random walks over the RDF graph to train a neural language model. *OWL2Vec** [3] is also a path-based approach but focuses on OWL ontologies instead of typical KGs to preserve the semantics of the graph structure, the lexical information and the logical constructors.
After generating the entities' embeddings for each SA, the cosine similarity between the vectors representing each entity in the pair corresponds to the graph embeddings similarity.

Supervised similarity learning tailored to similarity proxy To train a supervised semantic similarity according to the similarity proxy for which we want to tailor the similarity, eight representative ML algorithms for regression can be employed. *Linear Regression* and *Bayesian Ridge* assume there is a linear relationship between the independent and dependent variables. *K-Nearest Neighbor* explores the feature space and reaches a prediction for each sample based on the expected outputs of its neighbors. *Genetic Programming* is an evolutionary algorithm that tries to optimize a combination of variable and operators. *Decision Tree* predicts the value of a target variable by learning simple decision rules inferred from the data features. *Multi-layer Perception* is a class of feedforward artificial neural networks that learn non-linear functions through backpropagation of errors. *Random Forest* and *Extreme Gradient Boosting* (also known as XGBoost) are ensemble methods that combine the decisions from multiple decision trees to improve the overall performance.

These algorithms receive as input the semantic similarity values for the different SAs and the proxy similarity values as expected outputs. The output is an aggregated similarity score.

Supervised similarity evaluation The focus of the evaluation is to assess the ability of ML methods to learn combinations of SAs that improve the calculation of similarity. For each combination of an SSM with an ML algorithm, the Pearson's correlation coefficient is computed between the similarity proxies (expected values) and the obtained supervised similarity (predicted values). As baselines, our toolkit also computes the Pearson's correlation coefficient with the whole KG similarity, the single SA similarities and two well-known strategies for combining the single aspect scores (average and maximum).

3 Use case for the biomedical domain

Our toolkit was successfully applied in a set of protein and gene benchmark datasets [2], and two KGs including data from two biomedical ontologies, Gene

Ontology and Human Phenotype Ontology. These biomedical datasets rely on three proxies of similarity calculated based on mathematical expressions or other algorithms: protein function family similarity, protein sequence similarity and phenotype-based gene similarity. The results demonstrated our toolkit's ability to significantly produce semantic similarity models that fit different biological perspectives.

4 Conclusion

Our approach is independent of the SSM and the chosen ML method. Until now, we have used SSMs that take into consideration semantic and structural information. The inclusion of embedding methods that also consider lexical information should be incorporated into our toolkit in the future. In addition, although we only applied supervised ML algorithms to tailor semantic similarity to different biomedical similarity proxies, the proposed approach is versatile. As future work, we can evaluate our toolkit in other domain gold standards, such as the Lee50[2] where the similarity between news articles pairs has been been rated multiple times by humans and so it can be considered a similarity proxy.

Acknowledgements This work was funded by FCT through LASIGE Research Unit (UIDB/00408/2020, UIDP/00408/2020); projects GADgET (DSAIPA/DS/0022/2018) and BINDER (PTDC/CCI-INF/29168/2017); PhD grant SFRH/BD/145377/2019. It was also partially supported by the KATY project funded by European Union's Horizon 2020 research and innovation programme (GA 101017453).

References

1. Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Proc. of the 26th Int. Conf. on NIPS (2013)
2. Cardoso, C., Sousa, R.T., Köhler, S., Pesquita, C.: A Collection of Benchmark Data Sets for Knowledge Graph-based Similarity in the Biomedical Domain. Database **2020** (2020), baaa078
3. Chen, J., Hu, P., Jimenez-Ruiz, E., Holter, O.M., Antonyrajah, D., Horrocks, I.: Owl2vec*: Embedding of owl ontologies. Machine Learning pp. 1–33 (2021)
4. Pesquita, C., Faria, D., Bastos, H., Falcao, A., Couto, F.: Evaluating GO-based semantic similarity measures. In: Proc. of the 10th Annual Bio-Ontologies (2007)
5. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proc. of the 14th Int. Joint Conf. on AI (1995)
6. Ristoski, P., Paulheim, H.: RDF2Vec: RDF graph embeddings for data mining. In: Proc. of ISWC 2016 (2016)
7. Seco, N., Veale, T., Hayes, J.: An intrinsic information content metric for semantic similarity in wordnet. In: Proc. of the 16th European Conf. on AI (2004)
8. Sousa, R.T., Silva, S., Pesquita, C.: Evolving knowledge graph similarity for supervised learning in complex biomedical domains. BMC Bioinformatics (2020)
9. Yang, B., tau Yih, W., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases (2015)

[2] https://webfiles.uci.edu/mdlee/LeePincombeWelsh.zip