The Supervised Semantic Similarity Toolkit

Rita T. Sousa, Sara Silva, Catia Pesquita

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

See more details:

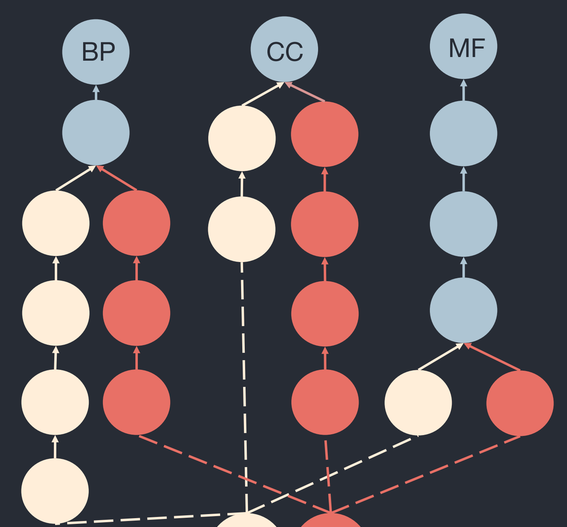
Motivation

Semantic similarity (SS) in knowledge graphs (KGs) is essential for several bioinformatics applications, namely protein-protein interaction prediction and disease-associated genes identification. Although KGs describe entities in terms of several semantic aspects, SS measures are general-purpose since they consider the whole graph.

This can represent a challenge since different use cases for the application of SS may need different similarity perspectives and ultimately depend on expert knowledge for manual fine-tuning.

1 Semantic aspects selection

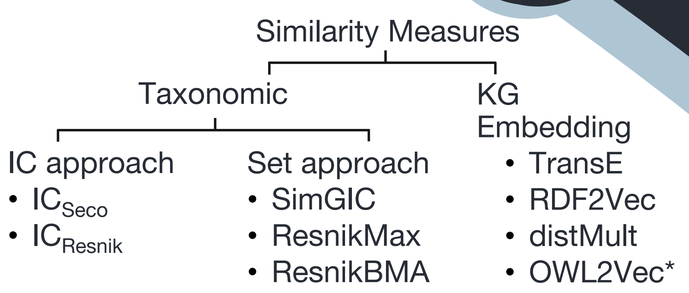
As default, our toolkit uses subgraphs rooted in the classes at a distance of one from the KG root class or the subgraphs when the KGs have multiple roots.



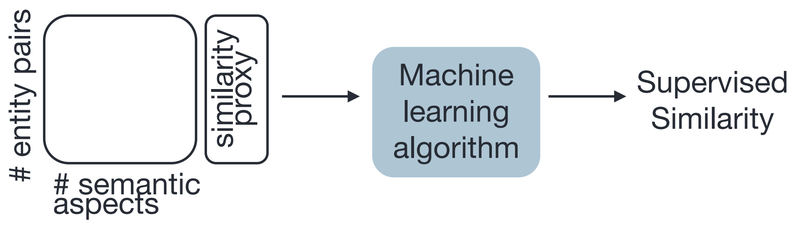
Gene Ontology (GO) describes protein function concerning three semantic aspects: the biological processes (BP) they intervene in, the cellular components (CC) where they are active and the molecular functions (MF) they perform.

2 Similarity computation for each semantic aspect

For the computation of KG-based similarities for each semantic aspect, the toolkit employs 10 SS measures: 4 based on KG embeddings and 6 based on taxonomic similarity.



3 Supervised similarity learning tailored to the similarity proxy



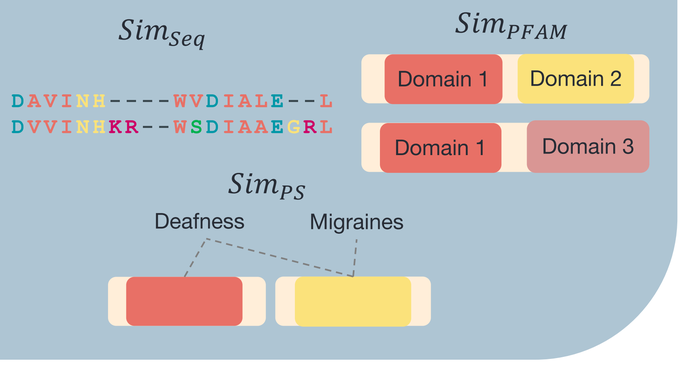
To train a supervised semantic similarity according to the similarity proxy, 8 representative ML algorithms for regression can be employed.

Use Case in the Biomedical Domain

The toolkit was successfully applied in a set of protein and gene benchmark datasets, and two KGs including data from Gene Ontology (GO) and Human Phenotype Ontology (HP).

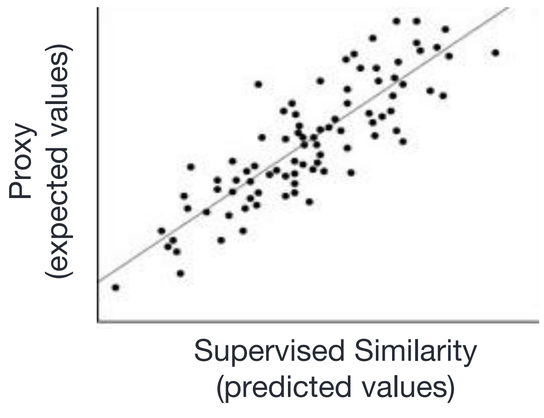
Dataset	Number datasets	Proxy	KG
Protein	10	Sim_{Seq}; Sim_{PFAM}	GO
Gene	1	Sim_{PS}	GO; HP

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.



4 Supervised similarity evaluation

As baselines, our toolkit computes the Pearson's correlation coefficient with the whole KG similarity, the single semantic aspect similarities and two well-known strategies for combining the single aspect scores.



Acknowledgments: 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 the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017453.