Published online: 2019-08-16

Appendix: Content Summaries of Best Papers for the 'Knowledge Representation and Management' section of the 2019 IMIA Yearbook

Arguello Casteleiro M, Demetriou G, Read W, Fernandez Prieto MJ, Maroto N, Maseda Fernandez D, Nenadic G, Klein J, Keane J, Stevens R

Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature

J Biomed Semantics 2018;9(1):13

This work combines semantic modelling (Cardiovascular Disease Ontology, CVDO) and learning algorithms (word embeddings). The authors aim at automatically identifying term variants or acceptable alternative free-text terms for gene and protein names from PubMed biomedical publications. Ontologies, such as CVDO, capture domain knowledge in a computational form and can provide context for gene/protein names as written in the literature. This study investigates: i) if word embeddings from Deep Learning algorithms can provide a list of term variants for a given gene/protein of interest; and ii) if biological knowledge from the CVDO can improve such a list without modifying the word embeddings created. The results are of significant performance improvements for deep learning algorithms on a gene/protein synonym detection task, by adding knowledge formalized in the CVDO (leveraging the formal relations between genes and proteins). Hence, the CVDO supplies the context that is effective to induce term variability for algorithms while reducing ambiguity. As a result, CVDO can be enriched with new discovered synonyms (*skos:altLabel*). This work relies on a generic approach to be reused with other medical ontologies.

Le KK, Whiteside MD, Hopkins JE, Gannon VPJ, Laing CR

Spfy: an integrated graph database for real-time prediction of bacterial phenotypes and downstream comparative analyses

Database (Oxford) 2018;2018:1-10

Spfy is a platform that rapidly performs the common reference laboratory tests owing to its database of diverse pre-computed results, and the ability to incorporate user data. This platform handles all analysis tasks by dividing them into subtasks, which are subsequently distributed across a built-in task management process. All results are converted into individual graphs and stored within a large graph database according to previously created ontologies: the Genomic Epidemiology Ontology (GenEpiO), the Feature Annotation Location Description Ontology (FALDO), and the Microbial Typing Ontology (TypOn). These ontologies provide the relevant metadata for genotypes, location, biomarkers, host, and source. In its presented version, Spfy contains 10,243 Escherichia Coli genomes, for which in-silico serotypes and Shiga-toxin subtypes, as well as the presence of known virulence factors and antimicrobial resistance determinants have been computed. Spfy includes analyses modules that are also self-contained and can be used in existing platforms. This work demonstrates that Spfy, by leveraging semantic technologies with a graph database, facilitates rapid phenotype identification, as well as the efficient storage and downstream comparative analysis of thousands of genome sequences.

Osumi-Sutherland DJ, Ponta E, Courtot M, Parkinson H, Badi L

Using OWL reasoning to support the generation of novel gene sets for enrichment analysis

J Biomed Semantics 2018;9(1):10

The Gene Ontology (GO) consists of over 40,000 terms for biological processes, cell components, and gene product activities linked into a graph structure by over 90,000 relationships. It has been used to annotate the functions and the cellular locations of gene products. The graph structure is used by a variety of tools to group annotated genes into sets whose products share function or location. These gene sets are widely used to interpret the results of genomics experiments by assessing which sets are significantly over- or under-represented in results lists. F Hoffmann-La Roche Ltd. has developed a manually maintained controlled vocabulary (RCV) for use in over-representation analysis. The formal structure of GO and logical queries in OWL allow to map RCV terms to sets of GO terms. Finally, gene sets derived from the resulting GO terms sets can be used to detect the signatures of cell and tissue types in whole genome expression data. This article is very interesting and demonstrates all the added-value of ontological representation to three axes: (i) it shows a practical use case of ontology-based reasoning and how the authors can solve problems with widely available standards and tools (OWL2 EL, ELK); (ii) in mapping from the RCV to the GO, the authors found and resolved over 200 omissions in the axiomatization; and (iii) the approach to automate mapping between RCV and GO, replacing the unsustainable manual mapping process.

Yu S, Ma Y, Gronsbell J, Cai T, Ananthakrishnan AN, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, Liao KP, Cai T

Enabling phenotypic big data with PheNorm

J Am Med Inform Assoc 2018;25(1):54-60

This paper addresses the difficulty to obtain a gold standard to train machine learning processes. The authors introduced a silver standard approach without human solicitation. They present PheNorm, a phenotyping algorithm that does not require expert-labeled samples at the training step. The input for the PheNorm algorithm consists of unlabeled data on a set of potentially informative features, either automatically curated or designed by experts. Online articles about the target phenotype from publicly available knowledge sources, such as Wikipedia and Medscape, are scanned with Natural Language Processing (NLP) software to extract medical concepts recorded in the Unified Medical Language System. These concepts are potentially related to the target phenotype. Then, narrative notes from the

Electronic Heath Record (EHR) database are processed with NLP software, which identifies mentions of the above medical concepts. With such a material, the most predictive features, such as the number of International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes or mentions of the target phenotype are normalized to resemble a normal mixture distribution with high area under the receiver operating curve (AUC) for prediction. The transformed features are then denoised and combined into a score for accurate disease classification. The authors validated the accuracy of PheNorm with four phenotypes: coronary artery disease, rheumatoid arthritis, Crohn's disease, and ulcerative colitis. The results suggest that PheNorm can potentially reduce the machine learning algorithm development process and demonstrate the capacity for EHR-driven annotations to scale to the next level – phenotypic big data.