PreMedOnto: A Computer Assisted Ontology for Precision Medicine

Noha S. Tawfik[1,2(✉)] and Marco R. Spruit[2]

[1] Computer Engineering Department, College of Engineering, Arab Academy for Science, Technology, and Maritime Transport (AAST), Alexandria 1029, Egypt
noha.abdelsalam@aast.edu

[2] Department of Information and Computing Sciences, Utrecht University, 3584 CC Utrecht, The Netherlands
{n.s.tawfik,m.r.spruit}@uu.nl

Abstract. This paper proposes an ontology learning framework that combines text mining, information extraction and retrieval. The proposed model takes advantage of existing structured knowledge by reusing terms and concepts from other ontologies. We further apply the methodology to create a detailed ontology for the emerging precision medicine (PM) domain by collecting a corpus of relevant articles and mapping its frequent terms to existing concepts. The resulting ontology consists of 543 annotated classes. The ontology was also tested for effectiveness by applying two evaluation frameworks to validate its design and quality. The results demonstrate that the ontology learning system is able to capture and represent the semantics of the PM domain with high precision and significance. Moreover, the computer-assisted construction process reduced dependency on expert knowledge. The developed *PreMedOnto* ontology could be further used to enhance the potentials of other NLP applications in the PM domain.

Keywords: Precision medicine · Data mining · Ontology reuse

1 Introduction

Ontologies are data models that transform domain's data into machine-readable representations to describe how a domain's information is organized. We adopt its original definition by Gruber as "An explicit specification of a conceptualization" [13]. By definition, they capture a wide variety of rich semantics by organizing knowledge into a hierarchy of concepts and relationships. It is considered one of the most reliable data representation models in today's semantic world, however, manual ontology development is an expensive task, both in terms of time and money. Ontology learning is the process of creating new ontologies from scratch whereas ontology population is concerned with augmenting existing ontologies with instances and properties. Both tasks require deploying efficient techniques to automatically process enormous amounts of domain-specific,

E. Métais et al. (Eds.): NLDB 2019, LNCS 11608, pp. 329–336, 2019.
https://doi.org/10.1007/978-3-030-23281-8_28

unstructured resources. While the latter task is hard, the former task is particularly challenging as computer models must closely mimic domain experts in interpreting meanings for constructing the ontology [7] and are usually accompanied by efficiency and precision issues. An alternative to overcome such limitations is to take advantage of existing knowledge bases, as not only it would minimize the human factor, but it would potentially achieve better precision and reduce redundancy [6]. Reusing contents would also guarantee a consistent representation of domain knowledge given the quality of the source ontology. The practice is quite established as part of the Web Ontology Language (OWL) specification and is also supported by the Open Biological and Biomedical Ontology (OBO) Foundry [17]. This study focuses on building an ontology for the precision medicine (PM) domain. The PM approach seeks to identify the best and the most effective practices for patients based on their genetic, environmental, and lifestyle factors. Although the concept has been around for many years, recently there has been an increase of public research funding and dedication to adopt the concept into practice versus the 'one-size-fits-all' method. Accordingly, there has been a substantial increase in the number of publications related to the PM concept [22]. However, the PM domain lacks a clear and organized hierarchy of its general, investigations, diagnostics and treatments' terminologies. The main contribution of this research is the compilation and development of the precision medicine ontology (*PreMedOnto*). Such an ontology helps in defining and shaping the precision medicine domain and its related vocabulary which improves the understanding of the field.

2 Related Work

In the recent years, ontology has become a preferable way to represent biological data [2]. There is a great amount of published research in the ontology engineering field, however, our survey is only limited to ontology engineering models built for the medical domain. Casteleiro et al. was able to build an ontology for the sepsis disease from an unannotated biomedical corpus. Their model used Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), as well as the neural language models Continuous Bag-of-Words (CBOW) and Skip-grams [5]. They also exploited the same model to enrich the cardiovascular diseases ontology (CVDO) from PubMed articles. A reuse-based method was proposed by Gedzelman et al. to construct another ontology for cardiovascular diseases [12] using UMLS and MeSH thesaurus. Cahyani and Wasito investigated the use of Ontology Design Patterns (ODP) to construct an Alzheimer's Disease ontology. Their model uses existing vocabulary and glossary to extract terms and relations from published articles and match them against the patterns [8]. Another Alzheimer's disease ontology was developed by Drame et al. [9], they cluster bilingual terms from English and French corpora, according to the UMLS thesaurus, and align them by integrating new concepts. In [16], the authors propose a framework for updating existing medical ontologies. Their approach consists of 4 steps: extract relevant terms, apply machine learning techniques to infer

polysemy, detect the concepts related to the term using clustering algorithms and finally, link terms to the exact positions in the ontology. Gao, Chen and Wang also suggested a model for extending ontologies [11] and applied it to the PHARE ontology. Their research took advantage of PMC repository to train a word2Vec model and uses random indexing to enrich ontology labels. In [15], Kang et al. attempted to tailor the general adverse event ontology to build specific diseases ontology (DSOAE). They used design patterns and addressed the specifications needed for the chronic kidney disease by adding new classes, relations and properties. Another model was proposed in [14], where the authors reused the existing GALEN ontology to build a specific ontology for the juvenile rheumatoid arthritis disease. Their semi-automatic approach relies on extracting relevant parts of the old ontology and refine them to ensure consistency and safety so that the semantics of imported concepts are not changed. Amato et al. [4] populated an ontology constructed by a domain expert with RDF templates extracted from medical records. Sanchez and Moreno [19] suggest a web based approach for building medical ontologies from scratch. It uses a set of user query words to collect web documents. Documents with the highest web search hit counts are considered valid taxonomic specialization for the domain. Named entities and verbs are then extracted to generate one-level taxonomy with general terms. The next stage is non-taxonomic learning where the extracted verbs are used as domain patterns and again used as web queries. Finally, the verb phrase is used to link each pair of concept. In [3], Alobaidi et al. combined UMLS thesaurus and Linked Open Data (LOD) classes to identify medical concepts and associate them to their corresponding formal semantics. Shah et al. constructed a framework based on MetaMap and SemRep to reuse terms from SNOMED-CT ontology. They applied the framework to construct an ontology that combines the dental and medical domain to allow better reasoning over common knowledge [20].

3 Methods

3.1 Proposed Model

Our ontology learning methodology is based on the concept of ontology reuse, where we adapt content from existing ontologies to model the PM domain. The model also relies on the assumption that the concepts that must be included in the ontology are mapped from the frequently mentioned terms present in the domain-specific data. And their co-occurrences frequency depicts the relations among them. To successfully achieve this goal, our proposed framework consists of 5 phases, Fig. 1 illustrates the overall learning process overview.

Knowledge Acquisition. In our work, we used a publicly available list of PM keywords and synonyms constructed by conducting a systematic search through multiple web resources, including: academic, news and health websites. As this list is manually compiled and verified, we refer to it as the PM vocab. The list is divided into three categories: keywords and synonyms for personalized medicine,

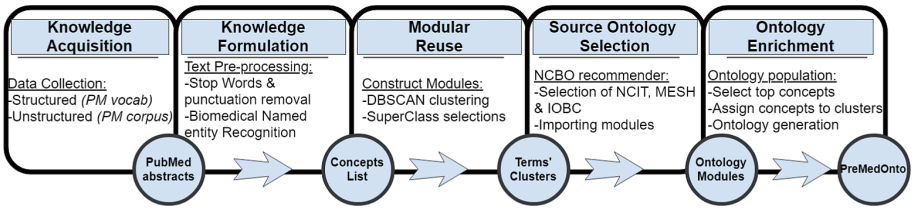

Fig. 1. Overview of the ontology learning framework.

keywords and synonyms for personal genomics and keywords and synonyms for diagnostics, biomarkers and testing. More details on the creation of the vocabulary could be found in [1]. In this paper we only use the last category since we aim at modelling the PM domain from a clinical and scientific point of view. In addition, we collected all titles and abstracts included in the PubMed repository discussing the PM concept. All articles included in PubMed are associated with Medical Subject Headings (MeSH) terms used for indexing articles. The search query used was "precision medicine" [Majr], adding the [Majr] term next to the original query restricts the search engine to return citations where the PM concept is the major focus of the article. In scientific literature, medical terminology is usually used interchangeably to describe the same concept. The MeSH entry terms or cross-references ensure that closely related terms and synonyms are all included when querying a certain term. In our case, the entry list has other terms such as Personalized Medicine and Individualized Medicine. The collection process was conducted through the Bio Python package that connects to NCBI E-utilities to retrieve and download articles. The results are then filtered so that all records with missing or incomplete abstract texts or in a foreign language other than English are excluded. This resulted in a total of 5,206 articles that serve as the *PM corpus*.

Knowledge Formulation. We preprocess all abstracts in the *PM corpus* to filter out stop words, symbols and punctuation. Due to the ambiguity of reporting biological or clinical results, MetaMap[1] was used for medical entity recognition. The output at this stage is a set of 6,832 distinct terms and concepts from the corpus. To guarantee precision, we do not map all terms extracted as this could lead to ambiguity and inconsistency in representing the domain knowledge. All terms mentioned more than once are ranked in descending order of their occurrence frequencies. Extracted terms are included only if their mention count exceeded a threshold. The threshold value is calculated as the weighted average occurrences of terms in documents to ensure that less significant words are removed.

Modular Reuse. In this stage, the *PM vocab* is used to create seed ontology modules where terms are mapped to a set of disjoint clusters. We started by analyzing the terms included in the *PM vocab* according to their relevance

[1] https://metamap.nlm.nih.gov/.

and commonness. We built a symmetric matrix of cosine similarity scores for every pair of word vectors that exist in the vocabulary. The word embeddings model was pretrained over a set of over 10 million biomedical articles from PubMed. The matrix was fed to a density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm implemented through the Scikit library. We opted for the DBSCAN clustering algorithm since it allows unsupervised learning over data and does not require the number of clusters a priori. This process created a total of 5 clusters. Following the creation of clusters, we rank all terms included according to their centrality and create one module per cluster. The top ranked concepts per cluster serve as the ontology super-classes. The original *PM vocab* set contained 100 terms that refer to diagnostic and testing procedures. Out of the 100 terms, only 73 were correctly clustered while 27 terms were regarded as noise by the clustering algorithm. Among the top candidate terms for each cluster, 25 were mapped as parent and child classes. Finally, we add all the non-used terms from the PM vocab to the list of concepts extracted from the *PM corpus*.

Source Ontology Selection. It is critical to determine the correct ontology that can serve as the base of the newly developed *PreMedOnto*. The criteria of choosing the ontology include coverage, acceptance and semantic language used. The NCBO ontology recommender is employed to suggest the best ontology for each module over all 895 existing ontologies. To maximize the coverage factor, we opted for the ontology set option which returns the best set of combined ontologies. The weights configuration for the recommender scoring function was set to the default settings. The final ranking of ontologies to be reused was: National Cancer Institute Thesaurus (NCIT)[2] , Medical Subject Headings (MeSH)[3] and Interlinking Ontology for Biological Concepts (IOBC)[4]. From the selected ontologies, we import all candidate classes with their ancestors, and verify that all remaining concepts per cluster are included in the module as child nodes. All redundant concepts in the *PM vocab* are removed by checking synonyms of each imported class.

Ontology Enrichment. In the final stage, each module is enriched by assigning relevant concepts extracted from the *PM corpus* in the knowledge formulation phase. We first extract the Uniform Resource Identifier (URI) corresponding to each concept. The ontofox [21] tool supports efficient ontology reuse by extending the Minimum Information to Reference an External Ontology Term MIREOT concept. The MIREOT approach favors selective class imports instead of importing the ontology as a whole. The ontofox web tool takes as input the base ontology, source terms URIs and parent classes URIs. It also allows users to choose the appropriate settings of the import process such as importing or omitting intermediate classes between input child and parent or deciding which annotation properties to return.

[2] https://bioportal.bioontology.org/ontologies/NCIT.
[3] https://bioportal.bioontology.org/ontologies/MESH.
[4] https://bioportal.bioontology.org/ontologies/IOBC.

3.2 Evaluation

Assessing the ontology output is a key factor in all ontology learning techniques. Not only to ensure the ontology quality before referencing and adopting it in other semantics-aware applications, but also to highlight errors and shortcomings. There are two different evaluative perspectives: ontology quality and ontology correctness. In this research, we carried out a two-fold evaluation process to measure the effectiveness of the constructed ontology: the first experiment assesses the ontology design whereas the second computes multiple quality features. To detect any design error in *PreMedOnto*, we use OntOlogy Pitfall Scanner (OOPS) online tool [18]. OOPS evaluates an OWL ontology against a catalogue of common mistakes in ontology The tool produces a summary of all pitfalls found within the ontology with extended information on each and a label indicating its importance level. We also apply the ontology quality evaluation framework (OQauRE) [10] to validate the quality of classes and axioms in *PreMedOnto*. OQauRe is a quantitative method based on the original software product quality requirements and evaluation concept. The framework computes multiple quality characteristics including structure, quality in use, reliability, compatibility, maintainability, operability, functional adequacy, transferability, performance efficiency. The generated metrics are mapped to quantitative values ranging from 1 to 5 with 3 is the minimum score and considered as accepted.

4 Results

The final output of the ontology learning process is the *PreMedOnto* in the standard OWL format. A total of 543 classes imported from 3 medical ontologies. Table 1 provides a brief summary of some of its metrics. The ontology can be accessed, viewed and downloaded from http://bioportal.bioontology.org/ontologies/PREMEDONTO.

Table 1. Summary of the *PreMedOnto* metrics generated by the Protégé framework.

Metric		Metric	
Classes	543	Classes with a single child	111
Average number of children	3	Maximum number of children	90
Properties	10	Maximum depth	7

The obtained results of evaluating *PreMedOnto* against the 41 pitfalls included in OOPS's catalogue, show that the ontology is free from critical and important pitfalls while there exist 3 cases of minor pitfalls. The former finding ensures the consistency and sustainability of the ontology, while the later might suggests corrections for better organization. The pitfalls detected are related to missing annotations, lack of connectivity and inverse relationship declaration. However, we find them irrelevant, as they do not threaten the functionality of the

ontology. The second experiment provides quantitative indicators of the quality of *PreMedOnto*. The computed scores for structure, compatibility and maintainability metrics were 3.5, 4.2 and 4.5 respectively. The ontology has successfully passed the minimal level required and is considered above average in most characteristics. It is worthy to mention that each quality measure is also associated to multiple sub-characteristics and hence indicates multiple quality aspects.

5 Conclusions

PreMedOnto is an application ontology built for the precision medicine domain on top of gold standard biomedical ontologies. The ontology learning process involves mining the PubMed repository to extract domain specific abstracts and vocabulary as sources of data. The information gathered is clustered and outlined to determine main modules. It reuses terms and concepts from NCIT, MeSH and IOBC to construct the ontology hierarchy. The evaluations demonstrate that the ontology content is reliable and consistent. We also plan to add a possible extra experiment to validate the ontology utility and applicability in the PM domain. The intended experiment involves human validation of the ontology by medical experts through a survey of questions.

References

1. Ali-Khan, S., Kowal, S., Luth, W., Gold, R., Bubela, T.: Terminology for personalized medicine: a systematic collection terminology for personalized medicine. Technical report (2016)
2. Alobaidi, M., Malik, K.M., Hussain, M.: Automated ontology generation framework powered by linked biomedical ontologies for disease-drug domain. Comput. Methods Programs Biomed. **165**, 117–128 (2018). https://doi.org/10.1016/j.cmpb.2018.08.010
3. Alobaidi, M., Malik, K.M., Sabra, S.: Linked open data-based framework for automatic biomedical ontology generation. BMC Bioinform. **19**(1), 319 (2018). https://doi.org/10.1186/s12859-018-2339-3
4. Amato, F., Santo, A.D., Moscato, V., Picariello, A., Serpico, D., Sperli, G.: A lexicon-grammar based methodology for ontology population for e-health applications. In: 2015 Ninth International Conference on Complex, Intelligent, and Software Intensive Systems. pp. 521–526. IEEE, July 2015. https://doi.org/10.1109/CISIS.2015.76
5. Arguello Casteleiro, M., et al.: Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature. J. Biomed. Semant. **9**(1), 13 (2018). https://doi.org/10.1186/s13326-018-0181-1
6. Bontas, E.P., Mochol, M., Tolksdorf, R.: Case Studies on Ontology Reuse. Technical report
7. Buitelaar, P., Cimiano, P., Magnini, B.: Ontology learning from text: methods. Eval. Appl. (2005). https://doi.org/10.1162/coli.2006.32.4.569
8. Cahyani, D.E., Wasito, I.: Automatic ontology construction using text corpora and ontology design patterns (ODPs) in Alzheimer's disease. Jurnal Ilmu Komputer dan Informasi **10**(2), 59 (2017). https://doi.org/10.21609/jiki.v10i2.374

9. Dramé, K., et al.: Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology: an application to Alzheimer's disease. J. Biomed. Inform. **48**, 171–182 (2014). https://doi.org/10.1016/J.JBI.2013.12.013
10. Duque-ramos, A., Duque-ramos, A., Fernández-breis, J.T., Stevens, R., Aussenac-gilles, N.: OQuaRE: a SQuaRE-based approach for evaluating the quality of ontologies. J. Res. Pract. Inf. Technol. **43**, 159 (2011)
11. Gao, M., Chen, F., Wang, R.: Improving Medical Ontology Based on Word Embedding (2018). https://doi.org/10.1145/3194480.3194490
12. Gedzelman, S., Simonet, M., Bernhard, D., Diallo, G., Palmer, P.: Building an ontology of cardio-vascular diseases for concept-based information retrieval. In: Computers in Cardiology, 2005, pp. 255–258. IEEE (2005). https://doi.org/10.1109/CIC.2005.1588085
13. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquisition **5**(2), 199–220 (1993). https://doi.org/10.1006/KNAC.1993.1008
14. Jiménez-Ruiz, E., Cuenca Grau, B., Sattler, U., Schneider, T., Berlanga, R.: Safe and Economic Re-Use of Ontologies: A Logic-Based Methodology and Tool Support. Technical report
15. Kang, Y., Fink, J.C., Doerfler, R., Zhou, L.: Disease specific ontology of adverse events: ontology extension and adaptation for chronic kidney disease. Comput. Biol. Med. **101**, 210–217 (2018). https://doi.org/10.1016/J.COMPBIOMED.2018.08.024
16. Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire, M.: A Way to Automatically Enrich Biomedical Ontologies. https://doi.org/10.5441/002/edbt.2016.82
17. Ochs, C., Perl, Y., Geller, J., Arabandi, S., Tudorache, T., Musen, M.A.: An empirical analysis of ontology reuse in BioPortal. J. Biomed. Inform. **71**, 165–177 (2017). https://doi.org/10.1016/J.JBI.2017.05.021
18. Poveda-Villalón, M., Carmen Suárez-Figueroa, M., Ángel García-Delgado, M., Gómez-Pérez, A.: OOPS! (OntOlogy Pitfall Scanner!): supporting ontology evaluation on-line. Technical report (2009)
19. Sánchez, D., Moreno, A.: Learning medical ontologies from the Web. Technical report
20. Shah, T., Rabhi, F., Ray, P., Taylor, K.: A guiding framework for ontology reuse in the biomedical domain. In: 2014 47th Hawaii International Conference on System Sciences, pp. 2878–2887. IEEE January 2014. https://doi.org/10.1109/HICSS.2014.360
21. Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A., He, Y.: OntoFox: web-based support for ontology reuse. BMC Res. Notes **3**(1), 175 (2010). https://doi.org/10.1186/1756-0500-3-175
22. Yates, L.R., et al.: The european society for medical oncology (ESMO) precision medicine glossary. Ann. Oncol. **29**(1), 30–35 (2018). https://doi.org/10.1093/annonc/mdx707