Computer Methods and Programs in Biomedicine 165 (2018) 117–128

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Automated ontology generation framework powered by linked biomedical ontologies for disease-drug domain

Mazen Alobaidi[a], Khalid Mahmood Malik[a,*], Maqbool Hussain[b]

[a] Department of Computer Science and Engineering, Oakland University, Rochester, MI, USA
[b] Department of Software, College of Electronics and Information Engineering, Sejong University, Seoul, South Korea

ARTICLE INFO

Article history:
Received 31 December 2017
Revised 31 July 2018
Accepted 14 August 2018

Keywords:
Semantic web
Ontology generation
Linked biomedical ontologies

ABSTRACT

Objective and background: The exponential growth of the unstructured data available in biomedical literature, and Electronic Health Record (EHR), requires powerful novel technologies and architectures to unlock the information hidden in the unstructured data. The success of smart healthcare applications such as clinical decision support systems, disease diagnosis systems, and healthcare management systems depends on knowledge that is understandable by machines to interpret and infer new knowledge from it. In this regard, ontological data models are expected to play a vital role to organize, integrate, and make informative inferences with the knowledge implicit in that unstructured data and represent the resultant knowledge in a form that machines can understand. However, constructing such models is challenging because they demand intensive labor, domain experts, and ontology engineers. Such requirements impose a limit on the scale or scope of ontological data models. We present a framework that will allow mitigating the time-intensity to build ontologies and achieve machine interoperability.

Methods: Empowered by linked biomedical ontologies, our proposed novel Automated Ontology Generation Framework consists of five major modules: a) Text Processing using compute on demand approach. b) Medical Semantic Annotation using N-Gram, ontology linking and classification algorithms, c) Relation Extraction using graph method and Syntactic Patterns, d), Semantic Enrichment using RDF mining, e) Domain Inference Engine to build the formal ontology.

Results: Quantitative evaluations show 84.78% recall, 53.35% precision, and 67.70% F-measure in terms of disease-drug concepts identification; 85.51% recall, 69.61% precision, and F-measure 76.74% with respect to taxonomic relation extraction; and 77.20% recall, 40.10% precision, and F-measure 52.78% with respect to biomedical non-taxonomic relation extraction.

Conclusion*:* We present an automated ontology generation framework that is empowered by Linked Biomedical Ontologies. This framework integrates various natural language processing, semantic enrichment, syntactic pattern, and graph algorithm based techniques. Moreover, it shows that using Linked Biomedical Ontologies enables a promising solution to the problem of automating the process of disease-drug ontology generation.

1. Introduction

There are growing pools of unstructured data relating to disease-drug interactions, and there is a huge potential for scientific insights and commercial innovation from using that unstructured data. For example, by using ontology/knowledge based information extraction [1] to build smart healthcare applications, machines can recognize both implicit and explicit entities in text [2]. These extracted entities, when used in machine learning algorithms in the next stage, help to improve the prediction accuracy of these machine learning algorithms [3]. However, this knowledge should be formally described in form of ontologies. On the other hand, unstructured data itself can be used to extract domain knowledge to build the ontologies. To turn that potential into valuable and beneficial information, a novel data model is needed to organize the disease-drug data and transform it into knowledge that is machine understandable. Recently, ontology has become the choice for representing and publishing biological knowledge. Ontology is defined as a formal representation of knowledge pertaining to a particular domain [4] or as defined by Gruber “explicit specification of a conceptualization” [5]. Ontology generation [6] is the process of building ontology for domains of interest by identifying the related concepts and the relations be-

[*] Corresponding author.
E-mail address: mahmood@oakland.edu (K.M. Malik).

https://doi.org/10.1016/j.cmpb.2018.08.010
0169-2607/© 2018 Elsevier B.V. All rights reserved.

Table 1
Advantages of extracting knowledge from LBO vs. various knowledge resources.

Knowledge resources	Publicly available	Machine understandable	Size	Reasoning	Biomedical domain	Approaches/Tools
Database schemas	Low	Low	Low	Low	Low	DB2OWL [10]: creating ontology from a relational database schemas
XML schemas	Low	Medium	Low	Low	Low	Syntactic Patterns
Wiktionary	High	Medium	High	Low	Low	OntoWiktionary [11]
LOD	High	High	High	High	Medium	mOntage [12]: ontology design and population framework
LBO + Unstructured Data	High	High	High	High	High	AOG-LBO

tween those concepts in those domains using (semi-) automated approaches. Apart from building the smart healthcare applications, ontologies and their knowledge-bases are fundamental cornerstones in the realization of semantic web vision [7]. In general, ontologies are built manually, but manual construction is labor intensive and requires domain experts and ontology engineers. Therefore, manually constructing ontologies on a large scale is not feasible, given its labor intensity. Therefore, this paper presents a framework called Automated Ontology Generation Framework Powered by Linked Biomedical Ontologies (AOG-LBO) that automates the process of generating Disease-Drug ontologies. It is governed by Linked Biomedical Ontologies (LBO) that aims to expose, share, and integrate related triples from diverse biomedical sources on the semantic web [8]. More specifically, in this work, we use Linked Life Data (LLD) [9], which is the first citizen of LBO, to generate Ontologies. Table 1, illustrates the advantages of using LBO compared to other resources to build ontologies. According to our best of knowledge, this is first attempt towards using LBO for automated ontology generation.

As illustrated in Fig. 1, we endeavor to inductively generate Disease-Drug ontology in a well-defined format from unstructured biomedical literature text by using LBO as clean knowledge constructed and verified by domain experts. Fig. 1 represents the AOG-LBO workflow, which illustrates the main processes along with their corresponding inputs (left side) and outputs (right side). As can be seen, there are three main processes in our ontology development framework. The first process is text processing (tokenization, segmentation, POS, and stemming). In turn the central part represents the medical semantic annotation, which aims to identify concepts in text, the semantic enrichment, which attempts to semantically enrich the concepts, and the relation extraction, which extracts relations between concepts. The last part represents the inference engine that aims to encode the semantic knowledge into formal ontological knowledge and infer new relations from existing encoded ontological knowledge. In this paper, we present a framework in which LBO is used as background knowledge beside biomedical text as input for automated ontology generation. The proposed framework leverages natural language processing (e.g. Part of Speech Tagging, n-gram, classification), semantic enrichment, and taxonomic relations to achieve a high degree of concept domain coverage.

Also, it uses graph method, and semantic enrichment to extract relations. Furthermore, the domain inference engine of the framework has the capability to generate well-defined ontology model. One advantage of our AOG-LBO framework, over recently developed frameworks, is its ability to discover only the relevant domain concepts from any given input text. For example, if the biomedical literature input is related to the protein domain, AGO-LBO will only discover the disease-drug concepts and ignore all other entities.

In summary, the main contributions of our proposed AOG-LBO Framework are as follows.

a) It incorporates LBO as potential knowledge source to construct well-defined ontologies
b) It automates the process of ontology extraction to support and facilitate the domain experts, and reduces the cost associated with ontology construction.
c) We implement novel medical semantic annotation and semantic enrichment algorithms to perform concept extraction using LBO as background knowledge.
d) It presents a novel taxonomic relation extraction algorithm by utilizing breadth first search and semantic enrichment while using LBO as a graph.
e) One of the salient feature of proposed framework is that it uses LBO, therefore, it doesn't require hand-crafted rules and training data.
f) It generates OWL full ontology that fulfills W3C semantic web standards.

2. Related works

There are four effective methodologies applicable to facilitate Ontology Generation that is related to our proposed approach, namely, rule-based & syntax analysis [13–19], syntactic pattern [20–25], machine learning [26–31], and Knowledge-based/External resources [32–37].

2.1. Rule-based & syntax analysis

The rule-based approach involves a manually crafted set of rules formed to represent knowledge that decides what to do or conclude across various scenarios. For example following set of simple grammar rules for person name recognition could be used: $< \text{salutation} > \langle \text{capsword} \rangle \ \langle \text{capsword} \rangle \rightarrow < \text{person} >$; where salutation includes "Mr. | Mrs. | Dr. | …", and capsword is a word that starts with caps letter. In [15] Abacha et al. focus on annotating medical entities and relationships from medical texts by using rule-based and syntactic patterns which are built semi-automatically from a corpus selected according to semantic criteria. In [16] Ono et al. proposed approach for extracting information on protein-protein interactions from the scientific literature based on lexicon, rule-based, and syntax analysis. In [17] Magka et al. proposed non-monotonic existential rules to axiomatize a variety of chemical classes based on functional groups. Typically, this approach achieves a very high level of precision, but quite low recall. This approach is labor intensive, works for one specific domain, and is less scalable [18,19].

2.2. Syntactic pattern

Syntactic pattern-based approach is a well-studied in the area of ontology engineering and has already proved to be successful in ontology extraction from unstructured text [20]. Unlike the rule-based approach, this approach comprises a large number of crafted syntactic patterns. Therefore, it has high recall and low precision. The crafted patterns are most likely broad and domain de-