Artificial Intelligence in Medicine 68 (2016) 47–57

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

A method for the development of disease-specific reference standards vocabularies from textual biomedical literature resources

Liqin Wang[a,b,*], Bruce E. Bray[a,c], Jianlin Shi[a], Guilherme Del Fiol[a], Peter J. Haug[a,b]

[a] Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Salt Lake City, UT 84108, USA
[b] Homer Warner Research Center, Intermountain Healthcare, 5121 South Cottonwood Street, Murray, UT 84107, USA
[c] Department of Internal Medicine, University of Utah, 30 North 1900 East, Salt Lake City, UT 84132, USA

ARTICLE INFO

Article history:
Received 5 June 2015
Received in revised form 22 February 2016
Accepted 25 February 2016

Keywords:
Knowledge extraction
Reference standards
Annotation
Saturation
Disease-specific ontology
Heart failure

ABSTRACT

Objective: Disease-specific vocabularies are fundamental to many knowledge-based intelligent systems and applications like text annotation, cohort selection, disease diagnostic modeling, and therapy recommendation. Reference standards are critical in the development and validation of automated methods for disease-specific vocabularies. The goal of the present study is to design and test a generalizable method for the development of vocabulary reference standards from expert-curated, disease-specific biomedical literature resources.

Methods: We formed disease-specific corpora from literature resources like textbooks, evidence-based synthesized online sources, clinical practice guidelines, and journal articles. Medical experts annotated and adjudicated disease-specific terms in four classes (i.e., *causes or risk factors*, *signs or symptoms*, *diagnostic tests or results*, and *treatment*). Annotations were mapped to UMLS concepts. We assessed source variation, the contribution of each source to build disease-specific vocabularies, the saturation of the vocabularies with respect to the number of used sources, and the generalizability of the method with different diseases.

Results: The study resulted in 2588 string-unique annotations for heart failure in four classes, and 193 and 425 respectively for pulmonary embolism and rheumatoid arthritis in *treatment* class. Approximately 80% of the annotations were mapped to UMLS concepts. The agreement among heart failure sources ranged between 0.28 and 0.46. The contribution of these sources to the final vocabulary ranged between 18% and 49%. With the sources explored, the heart failure vocabulary reached near saturation in all four classes with the inclusion of minimal six sources (or between four to seven sources if only counting terms occurred in two or more sources). It took fewer sources to reach near saturation for the other two diseases in terms of the treatment class.

Conclusions: We developed a method for the development of disease-specific reference vocabularies. Expert-curated biomedical literature resources are substantial for acquiring disease-specific medical knowledge. It is feasible to reach near saturation in a disease-specific vocabulary using a relatively small number of literature sources.

Published by Elsevier B.V.

1. Introduction

Disease-specific ontologies are knowledge bases intended to structure and represent disease-relevant information including disease etiology, diagnosis, treatment and prognosis. The availability of these ontologies could facilitate cross-disciplinary exchange and sharing of domain-specific knowledge. Disease-specific ontologies are also essential in supporting a variety of domain-specific computer applications, such as natural language processing, cohort selection, and clinical decision support [1,2]. For example, Haug et al. initiated a pneumonia-specific ontology, which supported the development of a clinical diagnostic modeling system [3]. Malhotra et al. constructed an Alzheimer's disease ontology and applied it to text mining on electronic health records [4].

However, the lack of comprehensive disease-specific ontologies hinders the development of such applications. BioPortal [5], an open repository of biomedical ontologies, currently hosts up to 467

* Corresponding author at: Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Salt Lake City, UT 84108, USA.
E-mail address: liqin.wang@utah.edu (L. Wang).

http://dx.doi.org/10.1016/j.artmed.2016.02.003
0933-3657/Published by Elsevier B.V.

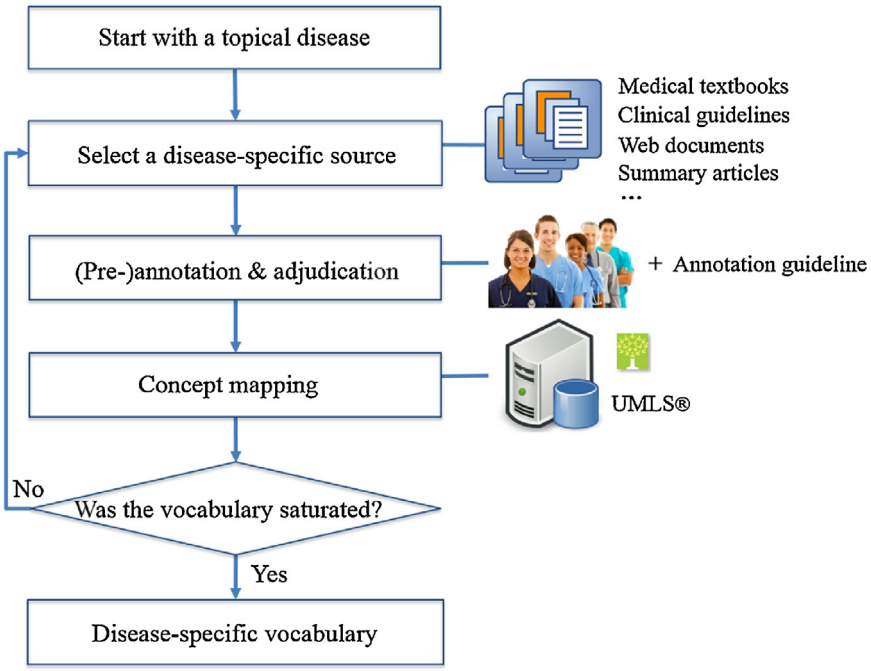

Fig. 1. Workflow for building near-saturated, disease-specific reference vocabularies from biomedical literature resources.

ontologies in various domains. However, among those ontologies less than 1% are disease-specific. Therefore, methods are needed to help develop disease-specific ontologies that can be made available to the community. A long-term goal of our research is to enable a platform that supports large-scale development of such ontologies.

Creating disease-specific ontologies is still a labor-intensive process. One of the main challenges is the knowledge acquisition, i.e., comprehensively ascertaining domain-specific concepts and relationships in the ontologies [6,7]. In knowledge engineering, domain experts are often used as the sources for acquiring medical knowledge. However, they are scarce and expensive. Another challenge is that while existing large terminologies, such as Disease Ontology [8] and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [9], can be used as sources of concepts for disease ontologies, the relationships between the concepts are primarily hierarchical, with little non-hierarchical relations between diseases and their signs and symptoms, diagnostic procedures, and treatments. Therefore, it is not feasible to extract a comprehensive set of disease-related relationships from those terminologies.

A promising alternative to address disease-ontology development challenges is to learn ontologies from textual data [7,10]. The learning can be separated into multiple levels: learning terms, synonyms, concepts, relations, axioms and rules [7,10]. At the term level, for instance, Riloff proposed a corpus-based approach for building domain-specific semantic lexicons [11]. At relationship level, Sanchez and Moreno studied methods that learn non-taxonomic relationships from web documents [12]. Particularly for developing disease-specific ontologies, the learning is primarily focused on using narrative text sources, such as the biomedical literature, to automatically identify disease-relevant concepts and relations. The relationships include the taxonomy backbone (i.e., is-a relations) and non-hierarchical relations (e.g., treats, causes). Most hierarchical relations between biomedical concepts are well represented in large domain ontologies and terminologies, such as SNOMED CT and the Unified Medical Language System (UMLS). However, important gaps still exist in regards to non-hierarchical relations. Learning these relations is an active subject of research interest [13–17].

The goal of the present study is to design and test a generalizable method for the development of vocabulary reference standards from expert-curated, domain-specific documents, such as textbooks, and clinical guidelines. The vocabularies and analyses established will be used to help the development and testing of automated disease-specific knowledge acquisition algorithms.

In the process of developing reference vocabularies, the number and types of sources that are needed to maximize the number of concepts retrieved are unknown. One source is unlikely to provide all concepts and relations about a disease and it is not feasible to manually extract concepts from all literature sources available. Therefore, in present study, we investigate the number of sources that are needed to obtain saturation for a disease-specific vocabulary. We assessed the feasibility of acquiring disease-specific concepts and relationships in the classes of *causes and risk factors*, *sign and symptoms*, *diagnostic tests and results*, and *treatments* by manually annotating terms from a representative and diverse set of popular knowledge sources in cardiology. Last, we then tested the generalizability of our methods with two additional diseases in *treatment* class.

2. Methods

In the present study, a disease-specific vocabulary is understood to be a list of concepts that are semantically related to a disease or syndrome. We focused on gathering disease-related concepts into a collection rather than identifying their taxonomic structure [18]. The framework for acquiring disease-specific vocabulary is displayed in Fig. 1. This is an iterative process with the goal of reaching near saturation which in this study is defined as finding <5% new concepts with the introduction of a new resource. We first formed a corpus with a collection of textual biomedical literature documents on the topical disease. Then, we initiated the iterative process by selecting one source from the corpus. The following step is annotation and adjudication, where the documents were annotated using eHOST, an open source annotation tool [15], by medical experts based on an annotation guideline. Any conflicted annotations were adjudicated by consensus between experts. Annotations were then mapped to equivalent UMLS concepts or, if these were unavailable,