Journal of Biomedical Informatics 63 (2016) 295–306

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Automated learning of domain taxonomies from text using background knowledge

Julia Hoxha [a], Guoqian Jiang [b], Chunhua Weng [a,*]

[a] Department of Biomedical Informatics, Columbia University, New York, NY, USA
[b] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

ARTICLE INFO

Article history:
Received 22 March 2016
Revised 18 July 2016
Accepted 1 September 2016
Available online 3 September 2016

Keywords:
Ontology learning
Taxonomy extraction from text
Semantic relation acquisition
Term recognition
Concept discovery

ABSTRACT

In this paper, we present an automated method for taxonomy learning, focusing on concept formation and hierarchical relation learning. To infer such relations, we partition the extracted concepts and group them into closely-related clusters using Hierarchical Agglomerative Clustering, informed by syntactic matching and semantic relatedness functions. We introduce a novel, unsupervised method for cluster detection based on automated dendrogram pruning, which is dynamic to each partition. We evaluate our approach with two different types of textual corpora, clinical trials descriptions and MEDLINE publication abstracts. The results of several experiments indicate that our method is superior to existing dynamic pruning and the state-of-art taxonomy learning methods. It yields higher concept coverage (95.75%) and higher accuracy of learned taxonomic relations (up to 0.71 average precision and 0.96 average recall).

1. Introduction

Ontologies are formal representations of knowledge resources that describe and share a common understanding of a particular domain. They are foundational for knowledge-based systems or intelligent systems and serve a wide range of applications such as Natural Language Processing [1], Information Retrieval [2], text clustering and classification, to name a few. Machine reading [3,4], which aims to extract structured knowledge from text with little human effort, has been a major goal of Artificial Intelligence since its early days and an important application area for ontologies. However, ontology development is a time and cost consuming task, requiring the knowledge of specialists from multiple disciplines who may have difficulties reaching consensus [5]. Current works in the field of automatic or semi-automatic ontology acquisition largely aim at overcoming this barrier.

Within this line of works, we present Ontofier, a novel framework to unsupervised ontology learning from text. In this work, we focus on the tasks of extracting domain concepts and their taxonomic relations. Concept hierarchies based on the taxonomic relations enable structuring information into categories, hence fostering efficient search, reuse, and formulation of relations.

In the wide spectrum of approaches to ontology classifications, as introduced in Uschold and Gruninger [6], at one end are the formal, heavyweight ontologies that make intensive use of axioms for specification, and at the other end are ontologies that use little or no axioms, referred to as lightweight ontologies. Taxonomies reside somewhere in the middle of this spectrum. Our contribution is a novel, fully-automated method for taxonomic relation learning from text. We present an extensive evaluation of our approach involving several medical experts, focusing on text from the biomedical domain, which is particularly challenging and lagging behind in ontology learning techniques. We used clinical trial eligibility criteria to illustrate our methodology, which promises to generalize beyond eligibility criteria text.

Potential applications of our approach include enrichment of current ontologies with new concepts and parent-child relations, improving text understandability for machines to allow better knowledge inference and search capabilities, and automated grouping of domain concepts for better engineering of classification features (e.g. Yu et al. [7] utilize a semi-automated approach for grouping of drug concepts to improve the classification features on drugs in phenotyping algorithms).

2. Related work

Ontology learning from text is the process of identifying terms, concepts, relations, and optionally axioms (for formal ontologies), from textual information and using them to construct and

* Corresponding author at: Department of Biomedical Informatics, Columbia University, 622 W 168th Street, PH-20, New York, NY 10032, USA.
E-mail address: chunhua@columbia.edu (C. Weng).

http://dx.doi.org/10.1016/j.jbi.2016.09.002
1532-0464/

maintain an ontology [8]. For our review, we consulted numerous surveys on ontology learning methods [8–12]. The learning techniques are generally categorized as *symbolic*, *statistical*, and *hybrid*. *Symbolic methods* rely on static linguistic patterns (rules) that can provide high accuracy, but require extensive domain expertise and are hard to generalize to other domains. Whereas *statistical methods* usually exploit corpora to learn structured knowledge, requiring minimal prior knowledge but providing better generalizability.

Our focus is on unsupervised statistical methods that do not require large amounts of labeled data. The most relevant works are based on clustering, which is useful for two purposes. First, similarity measures can provide information about the hierarchical relations of concepts. Second, the discovery of distinct clusters of similar terms can help to identify concepts and their synonyms. The works of [13–15] propose methods for unsupervised concept formation, whereas [16–19] introduce relation extraction techniques. These methods mainly make use of static, rare background knowledge.

In view of the shortcomings of conventional techniques, an interesting line of works is emerging. They explore the rich, heterogeneous resources of structured Web data for ontology learning. The intertwining of the Web with ontology learning enables us to harvest consensus (hence shared conceptualization) and access to large quantities of information. Among the few works [20–24] that explore structured Web data for relation extraction, Liu et al. [22] make use of Wikipedia's categorical system to deduce relations between concepts. They apply sentence parsers and syntactic rules to extract the explicit properties and values from the category names. Wong et al. [24] use Wikipedia and search engine page count to acquire coarse-grained relations between ambiguous concepts, using lexical simplification, and association inference. Mintz et al. [23] use Freebase as lookup dictionary to provide distant supervision for extracting relations between entity pairs. Fan and Friedman [25] introduced a distributional similarity approach for the semantic classification of concepts in the Unified Medical Language System (UMLS), the biggest repository of biomedical vocabularies.

As such, we observe an increasing trend in exploring structured web data for relation extraction. Boosheri and Luksch [26] also proposed an approach for ontology enrichment by using DBpedia. In contrast to our work, their approach extracts the relations (predicates) that DBPedia offers, relying first on a pre-defined similarity threshold to prune the predicates and then on ontology engineers to refine the recommended relations. Our work lies in the intersection between this framework of methods that use semantic knowledge bases in the Web, i.e. semantic-based techniques, and unsupervised statistical methods. This hybrid approach is relatively new and has not been well tested.

The novelty of our work lies in its exploitation of external knowledge bases in a fully-automated approach for concept formation and unsupervised taxonomical relation learning. In contrast to purely statistical methods, Ontofier employs not only text-based similarity measures but also concept semantic relatedness by using rich information of Web knowledge bases. Moreover, unlike symbolic methods, Ontofier does not rely on lexical patterns or rules manually crafted upon analysis of datasets/domain text at hand. Compared to existing clustering approaches, ours has unique advantages in dimensionality reduction and automatic clustering within each partition, not requiring pre-defined clustering parameters, which can be nontrivial and usually require costly fine-tuning procedures or prior expert knowledge.

3. Lightweight ontology learning

The commonality in various definitions is that an *ontology* is *a representation of entities* and *their relations* in *a particular domain* [9]. A key requirement is that each entity has *one unique reference*, an identifier, which is linked to *one or more natural language terms* to capture the synonymy inherent in human language. We adhere to this definition, using the following data structure for a domain concept.

Definition 1. A domain concept, extracted from a set $\mathcal{S}$ of natural language sentences of a particular domain, is defined as the tuple $c_i = (c_{id}, c_{name}, \mathcal{A})$, where c_{id} is a unique concept identifier, c_{name} is the concept name represented as a string, and $\mathcal{A}$ is the set of atoms composed of natural language phrases in the sentences $\mathcal{S}$ to which the concept is linked. Each atom is defined as $a_i = (a_{phrase}, s_l)$, s.t. a_{phrase} is the phrase (sentence fragment) linked to c_{id}, and $s_l \in \mathcal{S}$ the sentence where the phrase occurs.

Let us illustrate this definition with an example from real-life data on biomedical text in the domain of clinical trial patient recruitment.[1]

Example 1. The following text describes criteria of patients eligible in clinical trials for Alzheimer's disease:

"Exclude patients with a current diagnosis of hepatic or renal disease. Exclude patients with severe liver disorder or kidney disease."

We identify, among others, the domain concepts:

(c_1, *"liver disease"*, $\{a_1, a_2\}$) with atoms:

a_1 = (*"hepatic disease"*, s_1), a_2 = (*"liver disorder"*, s_2);

(c_2, *'kidney disease'*, $\{a_3, a_4\}$) with atoms:

a_3 = (*"renal disease"*, s_1), a_4 = (*"kidney disease"*, s_2);

The atoms represent the natural language terms to which a new concept is linked, also capturing the inherent synonymy. For brevity, we also use the concept notation $c_i = (c_{id}, c_{name})$, excluding atoms set.

An important piece of semantic information in an ontology is captured by the hierarchical relations among the concepts. According to formal, logic-based semantics, we are able to structure the ontology in the form of a hierarchy by determining subconcept/-superconcept relations (also referred to as subsumption relations) between the concepts [27].

According to the principles of subsumption theory, "to subsume is to incorporate new material into one's cognitive structures. When information is subsumed into the learner's cognitive structure it is organized hierarchically" [28]. Adhering to this theory, our learning process makes use of the *derivative subsumption*, which allows one to completely derive new concepts (as superconcepts) from an existing cognitive structure of known concepts. We use the following notation of the subsumption relation:

A subsumption relation, denoted as $\sqsubseteq_{(c_i,c_j)}$, is a binary relation of generic hierarchical nature between concept $c_i \in \mathcal{C}$ and concept $c_j \in \mathcal{C}$, where $c_i \sqsubseteq c_j$ states that the broader concept (or superconcept) c_j subsumes the more specific concept (or subconcept) c_i (i.e. $c_i \sqsubseteq c_j$). We can also state that c_i is subsumed by c_j.

Hence, the subsumption relation is used to create a hierarchy between general concepts and specific concepts. Referring to Example 1, by grouping the discovered concepts $(c_1, liver\_disease)$ and $(c_2, liver\_failure)$, we can derive a new concept $(p_1, DS\_liver\_disease)$ introducing the subsumption relations $\sqsubseteq (c_1, p_1)$ and $\sqsubseteq (c_2, p_1)$. Fig. 1 illustrates an excerpt of an automatically learned taxonomy in the biomedical domain.

[1] Text is extracted from the public portal http://www.clinicaltrials.gov.