Science of Computer Programming 95 (2014) 26–43

Contents lists available at ScienceDirect

Science of Computer Programming

www.elsevier.com/locate/scico

A domain-independent process for automatic ontology population from text

Carla Faria[a], Ivo Serra[b], Rosario Girardi[b,*]

[a] Computer Science Departament, Federal Institute for Education, Science, Tecnology of Maranhão (IFMA), São Luis, MA, Brazil
[b] Computer Science Departament, Federal University of Maranhão (UFMA), São Luis, MA, Brazil

HIGHLIGHTS

- We systematize the problem of Automatic Ontology Population.
- We propose a domain-independent process for Automatic Ontology Population.
- Our process overcomes the limitation of others of dependence of a domain.
- We conduct four experiments using a legal and a tourism corpora.
- Good effectiveness against its peers and adaptability are its main advantages.

ARTICLE INFO

Article history:
Received 1 February 2013
Received in revised form 2 September 2013
Accepted 2 December 2013
Available online 19 December 2013

Keywords:
Ontologies
Ontology population
Natural language processing
Information extraction

ABSTRACT

Ontology Population looks for instantiating the constituent elements of an ontology, like properties and non-taxonomic relationships. Manual population by domain experts and knowledge engineers is an expensive and time consuming task. Fast ontology population is critical for the success of knowledge-based applications. Thus, automatic or semi-automatic approaches are needed. This work proposes a generic process approaching the Automatic Ontology Population problem by specifying its phases and the techniques used to perform the activities on each phase. The main contribution of the work here described is a domain-independent process for the automatic population of ontologies from text that applies natural language processing and information extraction techniques to acquire and classify ontology instances. This is a new approach for automatic ontology population that uses an ontology to automatically generate rules to extract instances from text and classify them in ontology classes. These rules can be generated from ontologies of any domain, making the proposed process domain-independent and therefore, allowing the instantiation of ontologies quickly and at a low cost. Four experiments using a legal and a tourism corpora were conducted in order to evaluate the proposed process. Results indicate that this approach can extract and classify instances with high effectiveness with the additional advantage of domain independence. Some techniques representing the state of the art of this field are also described along with the solutions they adopt for each phase of the Automatic Ontology Population process with their advantages and limitations.

1. Introduction

Knowledge systems are a suitable computational approach to solve complex problems and to provide effective decision support. Their main components are a knowledge base and an inference mechanism to draw conclusions from that

[*] Corresponding author.
E-mail addresses: carlafaria@ifma.edu.br (C. Faria), ivocserra@gmail.com (I. Serra), rosariogirardi@gmail.com (R. Girardi).

0167-6423/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.scico.2013.12.005

knowledge. Knowledge representation formalisms, like ontologies, are used by modern knowledge systems, to represent and share the knowledge of an application domain [1].

Ontologies are an approach for knowledge representation capable of expressing a set of entities, their relationships, constraints and rules (conditional statements) of a given area [2,3]. These knowledge representation structures allow the semantic processing of information and, through more precise interpretation of data, systems have greater effectiveness and usability [4].

Knowledge acquisition is a costly and error prone process. Traditionally, ontologies are populated by domain experts and knowledge engineers, in a complex and slow task. This difficulty in capturing knowledge required by knowledge-based systems is known as the knowledge acquisition bottleneck. Therefore, it becomes crucial to automate this process.

Ontology Population (OP) looks for identifying instances of non-taxonomic relationships and properties of an ontology with knowledge discovered from different data sources such as text documents. Manual population by domain experts and knowledge engineers is an expensive and time-consuming task thus, automatic or semi-automatic approaches are needed. Performing a fast and low cost ontology population is crucial for success on the development of knowledge systems. OP occurs in both ontologies that have no instances and those already populated. When ontology population is performed on those that already have instances the process is known as Ontology Enrichment.

The OP problem is usually approached through the following three tasks: “Identification of Candidate Instances”, “Construction of a Classifier” and “Classification of Instances”. Each one of these tasks addresses a particular problem which is following discussed. In the “Identification of Candidate Instances” task the main challenges are both to work with texts in natural language and the identification of candidates instances. Texts in natural language are unstructured. Computer systems are able to understand instructions written in programming languages, but have difficulty to fully understand in natural language. This is due to the fact that programming languages are formal, containing fixed rules and well defined logical structures that allow computer systems to know exactly how to carry out each command. In natural language, a simple sentence usually contains ambiguities, nuances and interpretations depending on the context. So, the texts need to go through a previous structuring process, in order to identify candidates instances. Another challenge is the identification of named entities, unique objects that can be instances. In the “Construction of a Classifier” task, the main challenge is the automatic generation of a domain-independent classifier. A classifier can be constructed with either techniques from machine learning or information extraction. The generated classifier should be domain independent to reduce time and costs of ontology instantiation. With the application of techniques of supervised machine learning it is necessary to generate *n*-classifiers, one for each application domain from which we want to extract instances. The manual work for the construction of several classifiers makes it not feasible to apply supervised machine learning techniques. One solution would be the application of supervised machine learning techniques combined with unsupervised ones and information extraction techniques. In the “Classification of Instances” task the main challenge is the association of instances of properties and non-taxonomic relations to their respective classes with good effectiveness.

Most approaches for Automatic Ontology Population (AOP) [21,22,24,25,27–29,31,33,36] from texts are based on Natural Language Processing (NLP) [5,6], Statistic Models for Information Retrieval (SMIR) [7], Machine Learning (ML) [8] and/or Information Extraction (IE) [9,10]. NLP techniques are used to annotate the corpus with the information needed for subsequent processing. SMIR is used to extract from the corpus candidate instances. IE and ML techniques are used to validate the extracted instances and to classify them from the annotated corpus. One of the main limitations identified in the approaches is their dependence on a specific domain.

This paper discusses the problem of AOP and proposes a generic process to approach it specifying its phases and what kind of techniques can be used to perform the activities of each phase. Some techniques of the AOP state of the art are also described along with the solutions they adopt for each phase of the proposed AOP process.

This work, in contrast to related ones, proposes a domain-independent process to perform AOP from textual sources. This is a new approach for automatic ontology population that uses an ontology to automatically generate rules to extract instances from text and to classify them in ontology classes. These rules can be generated from ontologies of any domain, making the proposed process domain-independent. For evaluation purposes, four experiments were conducted in the legal and tourism domain, demonstrating its feasibility and effectiveness.

The article is organized in 7 sections following described. Section 2 introduces the formal ontology definition used in this work. Section 3 proposes a generic process approaching the OP problem along with the techniques that can be applied in each one of its phases. Section 4 gives an overview of the proposed process for AOP. Section 5 describes four experiments conducted to evaluate the proposed process. Section 6 summarizes related work and, finally, section 7 concludes the work.

2. A formal definition of an ontology

An ontology is a formal and explicit specification of a shared conceptualization of a domain of interest. Their classes, relationships, constraints and axioms define a common vocabulary to share knowledge [2]. Conceptualization refers to an abstract model of some phenomenon in the world. Explicit, means that the type of concepts used and the limitations of their use are explicitly defined. Formal, refers to the fact that the ontology should be machine readable. Shared, reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual but accepted by a group.

Formally, ontology can be defined as the tuple [4]:

$$O = (C, H, I, R, P, A)$$