Best Practice Poster:
Wikipedia-based Extraction of Lightweight Ontologies for Concept Level Annotation

Elshaimaa Ali
University of Louisiana at Lafayette, USA
eea7236@Louisiana.edu

Michael Lauruhn
Elsevier Labs, USA
m.lauruhn@elsevier.com

Keywords: Wikipedia; text mining; annotation; semantic annotation; lightweight ontologies

Abstract

This poster describes a project under development. We propose a framework for automating the construction of lightweight ontologies for semantic annotations. Lightweight ontology is defined as the ontology that does not have to include all the components expressed with formal languages such as concept taxonomies, formal axioms, disjoint and exhaustive decomposition of concepts. (Giunchiglia and Zaihrayeu 2009). However, manual enhancement of the ontology through the addition of axioms, rules, disjoint sets, etc., is possible for future reasoning purposes. The purpose behind this research is to evaluate possible means for efficiently annotating domain-specific content using open ontology sources.

When considering building ontologies for annotations in any domain, we follow the process of ontology learning in (Stelios 2006) which are: acquisition of the relevant terminology, identification of synonym terms / linguistic variants, formation of concepts, hierarchical organization of the concepts (concept hierarchy), learning of relations, properties or attributes, together with the appropriate domain and range, hierarchical organization of the relations (relation hierarchy), instantiation of axiom schemata, definition of arbitrary axioms, and ontology evaluation. Since we are looking for a lightweight ontology, we only consider a subset of these tasks, which are the acquisition of domain terminologies, generating concept hierarchies, learning relations and properties, and ontology evaluation.

When developing the framework modules we base most of our knowledge base on the structure of the Wikipedia, which represents the hierarchical links between categories and links between pages, in addition to specific sections of the content. To ensure machine readability and interoperability, ontologies have to be explicit to make an annotation publicly accessible, formal to make an annotation publicly agreeable, and unambiguous to make an annotation publicly identifiable (Ding 2006). An important aspect in order to achieve explicitly, formality and unambiguity of the developed ontology, is to define an annotation schema that allows the ontologies to be reused and be part of linked data.1 We designed our schema based on annotation elements already defined in the Dublin core standards2 and we used the DBpedia3 annotation elements for defining named entities. We are also introducing new elements for annotating concepts and defining the context (domain knowledge) in which the concept exists.

The main tasks for this framework are: extracting domain concepts and terms, measuring relatedness between domain terms, defining boundaries of subdomains using concept clustering and extracting relations, and defining named entities within each subdomain.

The following figure is an abstract explanation to the modules of the proposed framework.

1 http://linkeddata.org/
2 http://dublincore.org/documents/usageguide/
3 http://mappings.dbpedia.org/server/ontology/classes/
We start by defining a pool of domain terms and concepts that needs to be modeled for the domain ontology, but if this is not the case then we build Module A, where we start with relevant domain concepts and consider them as seed concepts for the ontology. Then we expand our concepts space using the Wikipedia link structure.

In Module B, we generate the domain and subdomain boundaries by computing relatedness between domain terms extracted in the first phase. Then we build a similarity matrix that models the relatedness between the extracted concepts. We developed a relatedness measure that relies on the degree of connectivity between two concepts in the Wikipedia graph. We then use hierarchical clustering (Diday and Simon 1980) to create subdomain boundaries. In Module C, we classify the generated named entities and concepts into wiki concepts and named wiki entities according to the description of the annotation schema. We will use the DBpedia triples for named entity recognition. In Module D we extract concept hierarchies and concept – concept relations by analyzing sections of Wikipedia articles. We will use openNLP to parse and extract relations defined in sections like the introductory sections in the Wikipedia page that defines the concept, in addition to exploring the category graph for the Wikipedia. OpenNLP has been successfully used for extracting relations for ontology enrichment in (Barkschat 2014). For Module E, we will evaluate the extracted ontology by comparing it to some of the mature existing ground truth like predefined domain ontologies or even topic maps that is created by a domain expert, and will use manual and expert evaluations.

**Acknowledgements**

Research and development of the project described in this poster and abstract is part of the LinkedUp Project, a European Commission FP7 Support Action.
References


