SEMCON:
A Semantic and Contextual Objective Metric for Enriching Domain Ontology Concepts

Zenun Kastrati, Norwegian University of Science and Technology, Gjøvik, Norway
Ali Shariq Imran, Norwegian University of Science and Technology, Gjøvik, Norway
Sule Yildirim-Yayilgan, Norwegian University of Science and Technology, Gjøvik, Norway

ABSTRACT
This paper presents a novel concept enrichment objective metric combining contextual and semantic information of terms extracted from the domain documents. The proposed metric is called SEMCON which stands for semantic and contextual objective metric. It employs a hybrid learning approach utilizing functionalities from statistical and linguistic ontology learning techniques. The metric also introduced for the first time two statistical features that have shown to improve the overall score ranking of highly relevant terms for concept enrichment. Subjective and objective experiments are conducted in various domains. Experimental results (F1) from computer domain show that SEMCON achieved better performance in contrast to tf*idf, $\chi^2$ and LSA methods, with 12.2%, 21.8%, and 24.5% improvement over them respectively. Additionally, an investigation into how much each of contextual and semantic components contributes to the overall task of concept enrichment is conducted and the obtained results suggest that a balanced weight gives the best performance.

KEYWORDS
Contextual Information, Domain Ontology Enrichment, Objective Metric, Ontology, Ontology Learning, Semantic Information, SEMCON

INTRODUCTION
Domain ontologies are a good starting point to model in a formal way the basic vocabulary of a given domain. They provide a broad coverage of concepts and their relationships within a particular domain. However, in-depth coverage of concepts is often not available, thereby limiting their use in specialized subdomain applications. It is also the business dynamics and changes in the operating environment which requires modification to an ontology (McGuinness, 2000). Therefore, the techniques for modifying ontologies, i.e. ontology enrichment, have emerged as an essential prerequisite for ontology-based applications. An ontology can be enriched with lexical data either by populating the ontology with lexical entries or by adding terms to ontology concepts. The former means updating the existing ontology with new concepts along with their ontological relations and types. This increases the size of the existing ontology which requires more computational resources and more time to compute. Thus making it less cost effective. The latter means adding new concepts without taking into account the ontological relations and types between concepts. As a result of this, the ontology structure will remain the same but its concepts will be enriched with their synonyms and homonyms.
Enrichment of ontology concepts aims to improve a given ontology by updating it with similar concepts. It is part of an iterative ontology engineering process (Faatz & Steinmetz, 2005) and it involves subtasks from only lower part of ontology learning layer cake model (Cimiano, 2006). Acquisition of the relevant terminology, identification of synonym terms or linguistic variants and the formation of concepts are subtasks involved. To perform these subtasks, the enrichment process requires an initial ontology which has to be enriched. It then explores available documents and texts from related domain of the given ontology in order to find synonyms or linguistic variants. Finally, by employing the learning approach, which is the core of an ontology concepts’ enrichment process, the concepts are ready for updating the initial given ontology.

There is a variety of learning approaches that are available to enrich concepts of an ontology. These approaches rely on either linguistic, pattern matching, machine learning or statistical techniques (Drumond & Girardi, 2008; Hazman, El-Beltagy, & Rafea, 2011). Even though these approaches have proved useful for enriching ontologies of many domains, they however have some limitations. These approaches use only contextual information without taking into account the semantic information of terms. The contextual information is derived by distributional property of terms such as term frequency or \( tf*idf \), and co-occurrence of terms. Therefore, to address this limitation, this paper proposes a new objective metric namely SEMCON to enriching the domain ontology with new concepts by combining contextual as well as semantics of a term.

The new proposed objective metric uses unstructured data as input for ontology learning process and is composed of two parts - contextual and semantic. Context is defined as the part of a text or statement – passage that surrounds a given term and it determines term meaning. In our work, it is the cosine distance between the feature vectors of any two terms. The feature vectors are composed of values computed by both the frequency of occurrence of terms in corresponding passages, and the statistical features such as font type and font size. The semantics on the other hand is defined by computing a semantic similarity score using lexical database WordNet.

In addition, we also have investigated into how much each of contextual and semantic components contributes to the overall task of enriching the domain ontology concepts and compared our results with the results obtained by other approaches such as \( tf*idf \), \( \chi^2 \) and LSA. We present our results for several domains, namely, Computer, Software Engineering, C++ Programming, Database and Internet.

The rest of the paper is organized as follows. Section Related Work presents the state of the art in the field of ontology enrichment. Section SEMCON describes our proposed SEMCON model in detail. In Section Experimental Procedures we describe the experiments including subjective and objective evaluation of SEMCON along with measures used to evaluate the effectiveness of objective methods. Results obtained by SEMCON and other objective methods and their comparisons are shown in Section Results and Analysis. Section Applications of SEMCON presents some of the application areas of SEMCON and lastly, Section Conclusion and Future Work concludes the paper and gives some future work directions.

RELATED WORK

The field of ontology learning from unstructured data has attracted a lot of attention recently, resulting in a wide variety of approaches. There are two main categories of these approaches relevant to the concept enrichment task: 1) Statistical approach, and 2) Linguistic approach.

Statistical approach uses distributional property of terms such as term frequency (tf) or term frequency inverse document frequency (tf*idf) and term co-occurrence to identify concepts within the textual data. An example of statistical approach as learning technique is DOODLE II (Yamaguchi, 2001). It exploits a machine readable dictionary and domain-specific texts to build domain ontologies with both taxonomic and non-taxonomic conceptual relationships. The non-taxonomic relationships are dependencies between concepts such as synonymy, meronymy, antonymy, attribute-of, possession.
Related Content

A Network Model Approach to Retrieval in the Semantic Web
[www.igi-global.com/article/network-model-approach-retrieval-semantic/2856?camid=4v1a](www.igi-global.com/article/network-model-approach-retrieval-semantic/2856?camid=4v1a)

Automatic Schema-Independent Linked Data Instance Matching System
[www.igi-global.com/article/automatic-schema-independent-linked-data-instance-matching-system/172424?camid=4v1a](www.igi-global.com/article/automatic-schema-independent-linked-data-instance-matching-system/172424?camid=4v1a)

OnGIS: Geospatial Data Integration Using Semantic Technologies and Query Containment
[www.igi-global.com/article/ongis/217010?camid=4v1a](www.igi-global.com/article/ongis/217010?camid=4v1a)
Semi-Automatic Ontology Construction by Exploiting Functional Dependencies and Association Rules
www.igi-global.com/article/semi-automatic-ontology-construction-exploiting/56465?camid=4v1a