knOWLearn: a reuse-based approach for building ontologies in a semi-automatic way

Samuel Vieyra[1,2], Mari Carmen Suárez-Figueroa[3], Hugo Estrada[1,2], Alicia Martínez[1]

[1] Departamento de Ciencias Computacionales, Centro Nacional de Investigación y Desarrollo Tecnológico (cenidet), Cuernavaca, Morelos, México.
{vieyra10c,hestrada,amartinez}@cenidet.edu.mx
[2]Fondo de Información y Documentación para la Industria INFOTEC, DF, México.
{samuel.vieyra,hugo.estrada}@infotec.com.mx
[3] Ontology Engineering Group (OEG), Facultad de Informática, Universidad Politécnica de Madrid, Campus Montegancedo, Boadilla del Monte, Madrid, España.
mcsuarez@fi.upm.es

Abstract. In this poster paper we present an overview of knOWLearn, a novel approach for building domain ontologies in a semi-automatic fashion.

Keywords: Domain ontology building, ontology learning, ontology reuse

1 Introduction

Ontologies are useful mechanism for representing knowledge, containing concepts and relationships about the domain of interest. Developing ontologies in a manual fashion is a complex and time consuming process, which implies the participation of domain experts and ontology engineers. For this reason, the definition of approaches to semi-automatically build domain ontologies, what is called ontology learning, is one of the main research topics in Ontology Engineering. However, current ontology learning approaches from textual documents have results not completely satisfactory [1]. In addition, such approaches do not consider the reuse of available domain ontologies, what implies (a) time-wasting while "reinventing the wheel" and (b) possible inclusion of errors in the semi-automatic ontology building. For these reasons, our approach for semi-automatically building domain ontologies from texts, called knOWLearn, is founded on the reuse of ontologies available on the web.

2 The knOWLearn Approach

Our approach to build domain ontologies consists of five main phases (see Fig. 1):
(1) ***Term Extraction***: this phase performs the extraction of relevant domain terms from text documents. FTC algorithm [2], to cluster documents, has been extended to obtain simple domain terms (of a single word). When the terms have been obtained, the most frequent n-grams (n={2,3}) containing such terms are searched in the input documents. These n-grams are the most relevant multi-words for the domain.

(2) ***Term Disambiguation***: a WordNet synset for each relevant term extracted in Phase 1 is obtained. For each term, all possible synsets are searched; and for each synset, the nouns that are found within the synset definition provided by WordNet are selected and matched with the input texts. Then, the synset that better matches with the input texts is selected. If none of the possible synsets has enough correspondences with the input texts, the term is considered as irrelevant to the domain.
(3) ***Ontology Building***: this phase builds an initial OWL ontology using the identified concepts (synsets). For this purpose, relevant domain ontologies are recovered with Watson[1] using the terms obtained in the two previous phases. In this way, previously defined knowledge is reused in the ontology building. Three different steps are carried out in this phase: (1) *search ontologies with Watson:* relevant terms are grouped in sets of three terms. This step permits recovering ontologies that contain relevant concepts and relations of a specific domain; (2) *finding mappings between ontology elements and Wordnet synsets:* hierarchical relations, synonyms, and string similarity metrics are used in this step; and (3) *ontology merging with WordNet* synsets: the obtained matches are joined in the initial ontology.
(4) ***Ontological Relations Inclusion***: this phase seeks ontological relations described within input texts. Patterns defined in [3] are used for finding these relations.
(5) ***Ontology Evaluation***: inconsistencies in the ontology are automatically detected and removed. In addition, a user can validate the content of the resulting ontology.

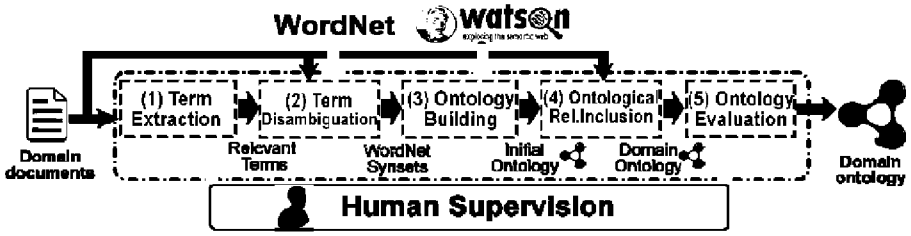

Fig. 1. General overview of knOWLearn

The proposed approach has been evaluated using three different domains: Breast Cancer, Enzyme Regulation and Molecular and Cellular Biology. The reuse of ontologies allowed us to obtain acceptable results in the accuracy of concept identification.

References

1. Zouaq A. and Nkambou R.: A Survey of Domain Ontology Engineering: Methods and Tools, in *Advances in Intelligent Tutoring Systems*, vol. 308, Springer, 2010, pp. 103–119.
2. Beil, F., Ester, M., Xu, X.: Frequent term-based text clustering. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 436–442.
3. Aguado de Cea, G., Gómez-Pérez, A., Montiel-Ponsoda, E., Suárez-Figueroa, M. C.: Using Linguistic Patterns to Enhance Ontology Development. International Conference on Knowledge Engineering and Ontology Development (KEOD), Madeira, Portugal, 2009.

[1] http:// watson.kmi.open.ac.uk /