A DOMAIN INDEPENDENT APPROACH FOR ONTOLOGY SEMANTIC ENRICHMENT

Tahar Guerram and Nacima Mellal

Departement of Mathematics and Computer Science,
University Larbi Ben M'hidi of Oum El Bouaghi - ALGERIA

ABSTRACT

Ontology automatic enrichment consists of adding automatically new concepts and/or new relations to an initial ontology built manually using a basic domain knowledge. In a concrete manner, enrichment is firstly, extracting concepts and relations from textual sources then putting them in their right emplacements in the initial ontology. However, the main issue in that process is how to preserve the coherence of the ontology after this operation. For this purpose, we consider the semantic aspect in the enrichment process by using similarity techniques between terms. Contrarily to other approaches, our approach is domain independent and the enrichment process is based on a semantic analysis. Another advantage of our approach is that it takes into account the two types of relations, taxonomic and non taxonomic ones.

KEYWORDS

Ontology, Ontology learning, Semantic enrichment, Natural language processing.

1. INTRODUCTION

According to AI community, ontology is a formal explicit specification of a shared conceptualization. [1]. Ontology enrichment is the task of extending an existing ontology with additional concepts and semantic relations and placing them at the correct position in the ontology [2]. Ontology learning is a wide research area that contains ontology enrichment, Ontology population and inconsistency resolution [2]. Ontology construction and maintaining, is a fastidious knowledge acquisition task which gives always a bottleneck problem, namely when the dynamicity of the ontology domain is high. In the other hand, because of the development of the world wide web, textual information is available with huge quantities. Hence, It will be very useful if this task is achieved automatically or semi automatically from textual sources. But automating ontology enrichment is not an end in itself but the objective is to preserve the coherence of the ontology after the enrichment process and the best way to achieve this is to consider the semantics of used texts.

In this paper, we propose an approach for semantic ontology enrichment. We begin by building an initial (or basic) ontology using a basic knowledge about a target domain. The semantic enrichment of this basic ontology is done through both syntactic and semantic analysis of a corpus of texts relating to the same target domain. Syntactic analysis is accomplished using natural language processing tools to obtain a POS tagged and named entity annotated corpus. We mention that before applying NLP tools, preprocessing operations of the studied corpus are

Natarajan Meghanathan et al. (Eds) : NLP, JSE, CST, SIP, ARIA - 2018
pp. 13– 19, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80202

applied like stop words eliminating and words stemming. For each sentence of this annotated corpus we extract a short sentence obeying to the form $<S_t\ V_t\ O_t>$ where S_t is the subject, V_t is the verb and C_t is the object. This short sentence plays the role of a relation $\mathbf{V_t(S_t, O_t)}$ which will be matched to the content of the basic ontology $< S_o\ V_o\ C_o >$. This matching is achieved using WorldNet resource by means of a semantic similarity measure and allows us to enrich the basic ontology as depicted in Figure 1 below.

The remainder of this paper is organized as follow. Section two is devoted to clarify the ontology learning bases. Section three gives a summary of related work; section four presents our approach. Finally, we conclude and we give some future work in section five..

2. ONTOLOGY LEARNING

In computer science, ontologies aim explaining and describing the world around us. However, in reality, they only focus on a part of the world, what is called a domain. The knowledge representation community defines ontology in accordance as follows: "Ontology is a formal, explicit specification of a shared conceptualization" [1]. 'Conceptualization' refers to an abstract model of phenomena in the world by having identified the relevant concepts of those phenomena. 'Explicit' means that the type of concepts used, and the constraints on their use are explicitly defined. 'Formal' refers to the fact that the ontology should be machine readable. 'Shared' reflects that ontology should capture consensual knowledge accepted by the communities." Simply, ontology represents the knowledge by a set of the concepts within a domain of interest and the relationships between those concepts. For that aim, ontologies play a central role in knowledge extraction, they can be learnt from various sources, be it databases, structured and unstructured documents or even existing preliminaries like dictionaries, taxonomies and directories.

To a large extent, the ontology learning system is understood in a variety of ways, it can be ontology extraction, ontology generation, or ontology acquisition. Nevertheless, ontology learning can be defined as an automatic or semi-automatic creation of ontologies, including extracting the corresponding domain's terms and the relationships between the concepts that these terms represent from a corpus of natural language texts, and encoding them with an ontology language for easy retrieval.

Ontology enrichment is one of the important objectives for the ontology learning process. It consists of adding automatically new concepts and new relations to an initial ontology constructed manually using a basic knowledge relating to a given domain. Concepts and relations have to be placed in the relevant place in the initial ontology. However, numerous approaches and applications focus only on constructing taxonomic relationships (is-a-related concept hierarchies) rather than full-fledged formal ontologies [3]**.** For that, we are interesting, in our work to develop an approach for the ontology enrichment taking in account both taxonomic and non-taxonomic relationships between concepts. Generally, the process of enrichment attempts to facilitate text understanding and automatic processing of textual resources, moving from words to concepts and relationships. It starts by extracting concepts/relationships from plain text using linguistic processing such as part-of-speech (POS) tagging and phrase chunking [4]. The extracted concepts and relationships are then arranged in the initial ontology, using syntactic and semantic analysis techniques.

3. RELATED WORK

The first paper [5] presents a methodology called PACTOLE (Proprietary And Class Characterization from Text for Ontology Enrichment) for the enrichment of an initial ontology

from a collection of texts relating to the astronomic domain. The first step is analyzing the collection of texts using NLP techniques in order to extract objects of the domain and their properties using predefined syntactic patterns then in the second step FCA technique is applied to the couples (object, property) in order to generate a concept lattice where each concept is a collection of a maximum number of objects sharing the maximum number of properties. The third step consists of expressing existing celestial objects data base by a second lattice of concepts using FCA technique as well as. The fourth step consists of merging the two lattices of concepts to obtain a resulting concept hierarchy. In step five this concept hierarchy is represented in FLE description language to be able to do reasoning tasks on it. The methodology was applied on a high number of Astronomy Abstracts journals and with the existing SIMBAD celestial objects database and the score of precision is high (74.71%) meaning that objects are classified in adequate classes and the score of recall is low because, mainly, the number of properties associated with objects is not sufficient for classification.

The second paper [6] presents a framework called Ontorich allowing the enrichment and the evaluation of ontologies using RSS Feeds. The enrichment of an otology is proceeded using OpenNLP API, which is a natural language processing Library, and WordNet [7] resource. RSS feeds are an important source of information as they provide permanently updated web information. To extract relations and concepts from RSS feeds, statistical and syntactic methods are applied using OpenNLP API. After the enrichment phase, the author(s) used several metrics to measure how the initial ontology is modified. Ontorich was compared against two ontology enrichment systems, which are Kaon and Neaon, and compared also against two other ontology evaluation systems, which are OntoQA and Romeo, relating to a certain number of functionality criteria and the results show that Ontorich is a more powerful tool for ontology enrichment and evaluation.

The third paper [8] presents a framework based on machine learning strategy to the automatic extraction of non taxonomic relations which remain a great challenge for ontology learning systems community. The framework proposed, initially extracts a set of causation contextual constructs (CCC) from annotated corpus and WordNet [7] to be used as initial indicators that can locate the good candidate sentences that may hold causation relation in text. In the second step, a new algorithm (graph based semantics GBS) is applied to indicate the real existence of causation in the sentences and if so, label both relation parts (cause, effect). To achieve this, sentences are divided into two parts and the most representative word in each part is searched based on the hypernym structure. The main steps of this algorithm are:

1. Specify causation relations direction according to CCC (cause-effect or effect-cause).

2. Extract the window for each relation part (for both cause and effect).

3. Build a graph for each window.

4. Specify the RDB (relation data base containing examples of cause, effect) semantic pair, suitable for the window.

5. Process each window graph to find best candidate semantic feature from the graph.

6. Extract a representative noun in the window that corresponds to the semantic feature.

To evaluate their system, precision, recall, and F-measure are computed based on a set of a total of 1213 used sentences and the results were as follow: precision = 78 %, recall = 68% and F-measure = 73 %.

The fourth paper [9] proposes an automatic process for ontology population from a corpus of texts. The proposed process is independent from the domain of discourse and aims to enrich the initial ontology with non taxonomic relations and ontology class properties instances. This process is composed of three phases: identification of candidate instances, construction of a classifier and classification of the candidate instances in the ontology. The "Identification of instance candidates phase applies natural language processing techniques to identify instances of non taxonomic relationships and properties of an ontology by annotating the inputted corpus. The "Construction of a Classifier" phase applies information extraction techniques to build a classifier based on a set of linguistic rules from ontology and queries on a lexical database. This phase has a corpus and an ontology as inputs and outputs a classifier used in the "Classification of Instances" phase to associate the extracted instances with ontology classes. Using this classifier, an annotated corpus and the initial ontology, the third phase consisting of the classification of these instances, produces a populated ontology.

Implementation of this process applied to the legal domain show results of 90% as precision 89.50% as Recall and 89.74% as F-measure. Authors conducted others experiments of their process on the touristic domain and obtained the results of 76.50% as precision 77.50% as Recall and 76.90% as F-measure.

The fifth and the last, but not least, paper [10] presents a pattern based approach of ontology enrichment by antonymic relations extracted from Arabic language corpora. Ontology of "seed" pairs of antonyms is used to extract lexicon - syntactic patterns in which pairs of antonyms occur. These patterns are then used to find new antonym pairs in a set of Arabic language corpora. The approach is tested on three different Arabic corpora: classical Arabic corpus (KSUCCA) [11], the contemporary Arabic corpus (CCA) [12] and the mixed Arabic corpus (KACSTAC) [13]. The correctly extracted patterns are used to enrich an ontology based lexicon for Arabic semantic relations called SemTree [14]. The developed system has as input the set of patterns and the KSUCCA corpus. First the given corpus is preprocessed in order to clean diacritics from the texts and by pattern matching antonyms are extracted and evaluated by an expert evaluator and new antonym pairs are added to the SemTree ontology. The system is evaluated using three measures which are pattern reliability, precision and performance of the system. Pattern reliability is the ratio of correct antonyms extracted using the pattern to the total extracted ones using the same pattern. System precision is the ration of the total correct extracted antonyms to the total extracted ones, while system performance is the measure of the increase in ontology size. The obtained results show that despite the fact that system performance is high (42, 3 %), system precision computed is about 29, 45 % as a mean of all obtained precisions relating to all used corpuses (KSUCCA, CCA and KACSTAC).

To summarize, we can say that firstly, all the above approaches consider only one type of relations, taxonomic or non taxonomic. Secondly, the performance of the above approaches depends on the target domain.

Our proposed approach aims to consider the two kinds of relations, taxonomic or non taxonomic, and to preserving the coherence of the enriched ontology by using the semantic similarity measure techniques offered by WordNet [7] technology, and this, independently of the domain of discourse.

4. THE PROPOSED APPROACH

We propose an approach for automatic ontology enrichment giving a corpus of texts relating to a target domain. First, a basic knowledge related to this target domain is predefined and represented in an initial or a basic ontology through a set of concepts and relationships between

these concepts. The objective is to enrich the initial ontology by the content of texts relating to the same target domain through semantic analysis.

The proposed enrichment process is composed of three phases. In the first phase, we proceed to the annotation of the texts using a morpho-syntactic analysis of the given texts by means of natural language processing tools in order to provide a first level of understanding of the given texts. We parse texts to extract syntactic relations between terms as well as the part of speech tags of these terms. This annotation phase is followed by a second phase consisting of a simplification of complex sentences to simple clauses. It consists of a semantic analysis of the annotated text where clauses obeying to the form V(S, O) (S: subject, V: verb, O: object). The Third phase called semantic enrichment phase, consists of the comparison of each extracted relationship to the content of the initial ontology using a similarity measure. According to this comparison, we decide whether each extracted relationship will be candidate to enrich our initial ontology or not. The similarity measure is based on WordNet [7] and the enrichment process aims to identify new concepts or new relationships or just concept or relationship instances already existing. In this step, we study for each relation $V_t(S_t, O_t)$, extracted from the text, the semantic similarity of this last with existing ontology relations $V_o(S_o, O_o)$ to identify new concepts and relations enriching the ontology. Figure 1 gives the semantic enrichment algorithm of the basic ontology (Phase 3 of the semantic enrichment framework given by the figure 2). In figure 1, *SimilarityThreshold* is a parameter of the algorithm fixed by the user or by the domain expert.

Case 1 : Semantic_Similarity (V_t ,V_o) **> =** SimilarityThreshold , **THEN**
*/ * study of the similarity between S_t & S_o, and between O_t & O_o, */*
1- **IF** S_o & S_t are locally linked in the ontology, **THEN** do nothing.
ELSE use **WordNet** technique to extract this link between S_o & S_t ;
IF link exists, then add it to the ontology, **Else,** add S_t as a concept (class) which will be the domain of V_t ;
2- **IF** S_o & S_t are locally linked in the ontology, **THEN** do nothing.
ELSE use **WordNet** to extract this link between S_o & S_t ;
IF link exists then add it to the ontology, **Else** add S_t as a class, which plays the role of V_t domain;
3- **IF** O_t & O_o are locally linked in the ontology, **THEN** do nothing.
ELSE use **WordNet** to extract this link between O_t et O_o;
IF link exists then add it to the ontology, **Else** add O_t as a class, which plays the role of V_t Codomain ;

Case 2 : Semantic_Similarity (V_t ,V_o) < SimilarityThreshold **THEN**
add $\mathbf{V_t}$ to the ontology as a relation and using **WordNet** find **Link(V_t ,V_o).**
IF Link (V_t , V_o) found then add it to the ontology
*/ * study the similarity between S_t & S_o, and between O_t & O_o, */*
: 1- **IF** Semantic_Similarity (S_t , S_o) **> =** SimilarityThreshold (same appellation), **THEN** define S_o as domain of V_t .**ELSE** use **Wordnet** to find Link (S_t, S_o), **IF** Link**(S_t, S_o)** found **THEN** add it to the ontology, **ELSE** add S_t as Class in the ontology (represents the domain of V_t).
2- **IF** Semantic_Similarity(O_t, O_o) **>=** SimilarityThreshold (same appellation), **THEN** define O_o as codomain of $\mathbf{V_t}$. **ELSE** using **Wordnet** find Link(O_t, O_o), **IF** Link(O_t, O_o) found **THEN** add it to the ontology, **ELSE** add O_t as a Class in the ontology (represents the codomain of V_t).

Figure 1: Semantic ontology enrichment

Our objective is to get an enriched ontology giving an extended semantic coverage of a target domain. We give below in Figure 2, the semantic enrichment framework of the proposed approach.

5. CONCLUSION

In this paper we have proposed an approach for ontology enrichment. It is composed of three phases. The first phase consists of the annotation of the corpus of texts relating to a given domain using natural language processing tools. The second phase allows extracting knowledge from the annotated corpus of texts in the form of basic binary relations. The third phase consists of the semantic enrichment of the basic ontology using WorldNet similarity techniques. Besides the consideration of all types of relations, our approach presents two main advantages, compared to the existing approaches. The fist advantage of our approach is that it is independent from the domain of discourse and the second one is that the enrichment process is done using semantic similarity between relations and concepts which allows preserving the coherence of the enriched ontology. Actually, we are building a basic ontology relating to Small and Medium sized Enterprises (S.M.E) domain in the aim to validate our approach and we expect to obtain promoting results.

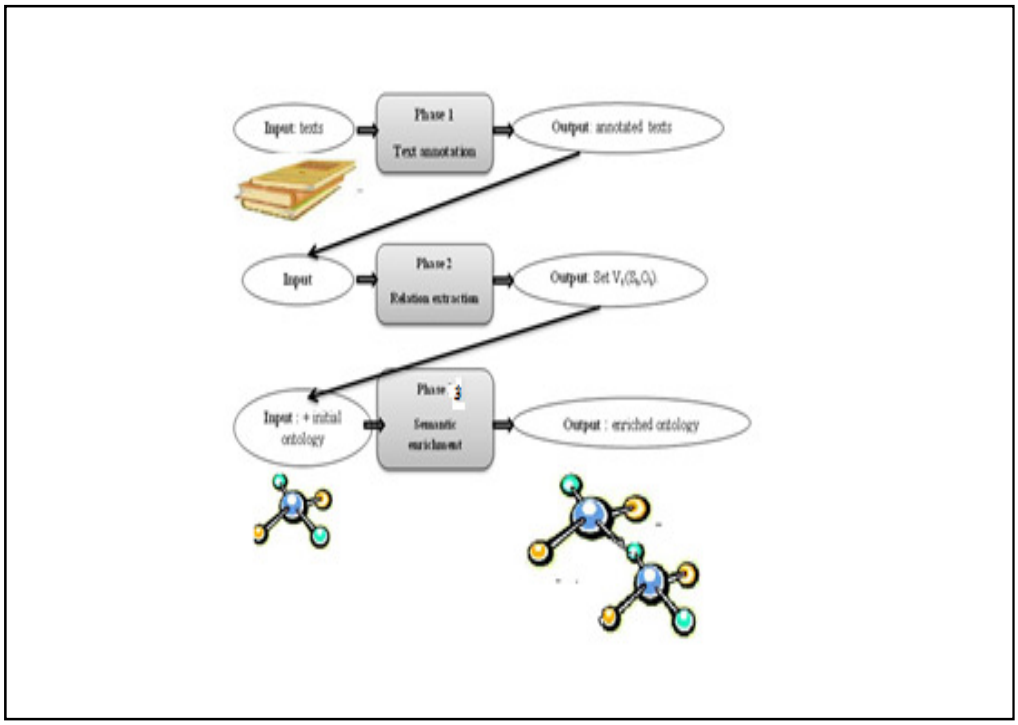

Figure 2: The semantic enrichment framework of the proposed approach

REFERENCES

[1] T. R. Gruber, "A translation approach to portable ontology specifications," Knowledge acquisition, vol. 5, no. 2, pp. 199-220, 1993.

[2] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, and E. Zavitsanos, "Ontology population and enrichment: State of the art, " In Knowledge-driven multimedia information extraction and ontology evolution, Springer-Verlag, January 2011, pp. 134-166

[3] C. Biemann, " Ontology learning from text: A survey of methods,". In LDV forum , Journal of Computational Linguistics and Language Technology, vol. 20, no. 2, pp. 75-93. 2005

[4] A. Gómez-Pérez, and D. Manzano-Macho,. "An overview of methods and tools for ontology learning from texts," The knowledge engineering review, vol. 19, no. 3, pp. 187-212, 2004

[5] R. Bendaoud, Y. Toussaint and A . Napoli, " Pactole: A methodology and a system for semi-automatically enriching an ontology from a collection of texts," Lecture Notes in Computer Science, vol. 5113, pp. 203-216, 2008.

[6] G. Barbur,., B. Blaga, and A. Groza, " OntoRich; A support tool for semi-automatic ontology enrichment and evaluation," In IEEE International Conference on Intelligent Computer Communication and Processing, 2011, pp. 129-132,

[7] Princeton University, "WordNet : A lexical Data Base for English" , Wordnet Princeton University , 2010. [online]. Available: http://wordnet.princeton.edu/wordnet, [Accessed : June, 10 th , 2016]

[8] A. S. Al Hashimy and N. Kulathuramaiyer, "Ontology enrichment with causation relations, " In IEEE Conference on Systems, Process & Control (TCSPC 2013), 2013, pp. 186-192.

[9] C. Faria, I, Serra, and R. Girardi, "A domain-independent process for automatic ontology population from text," Science of Computer Programming, vol. 95, pp. 26-43, 2014.

[10] M. Al-Yahya, S. Al-Malak, and L. Aldhubayi , "Ontological lexicon enrichment: The BADEA system for semi- automated extraction of antonymy relations from Arabic language corpora," Malaysian Journal of Computer Science. vol. 29, no. 1, 2016.

[11] Classical Arabic corpus (KSUCCA), [online]. Available: http://www. ksucorpus.ksu.edu.sa. [Accessed: july, 13, 2017] .

[12] Contemporary Arabic corpus (CAC) . [Online]. Available: http://www.comp.leeds. ac.uk/eric/ latifa/ research.htm . [Accessed: july, 13, 2017] .

[13] Mixed Arabic corpus (KACSTAC) . [online]. Available : http://www.kacstac.org.sa/pages / Default.aspx. [Accessed: july 13, 2017] .

[14] A. Al-Zahrani,., Al-Dalbahie, M., Al-Shaman, M., Al-Otaiby, N., and W. Al-Sultan,. SemTree: analyzing Arabic language text for semantic relations. PhD Thesis, IT Department, KSU, Saudi Arabia,, 2012.

AUTHORS

Tahar Guerram is assistant professor of Computer Science at the Department of Mathematics and Computer Science at the university of Larbi Ben Mhidi of Oum El Bouaghi and head of the same department. His areas of interest include mutli agent systems, text mining, ontologies and natural language processing.

Nassima Mellal is assistant professor of computer science at the Department of Mathematics and Computer Science at the university of Larbi Ben Mhidi of Oum El Bouaghi. Her areas of interest include semantic web and e-learning.