Effective Ontology Learning :
Concepts' Hierarchy Building using Plain Text Wikipedia

Khalida Ben Sidi Ahmed, Adil Toumouh, and Mimoun Malki

Department of Computer Science, Djillali Liabes University,
Sidi Bel Abbes, Algeria
send.to.khalida@gmail.com

Abstract. Ontologies stand in the heart of the Semantic Web. Nevertheless, heavyweight or formal ontologies' engineering is being commonly judged to be a tough exercise which requires time and heavy costs. Ontology Learning is thus a solution for this exigency and an approach for the 'knowledge acquisition bottleneck'. Since texts are massively available everywhere, making up of experts' knowledge and their know-how, it is of great value to capture the knowledge existing within such texts. Our approach is thus an interesting research work which tries to answer the challenge of creating concepts' hierarchies from textual data. The significance of such a solution stems from the idea by which we take advantage of the Wikipedia encyclopedia to achieve some good quality results.

Keywords : domain ontologies, ontology learning from texts, concepts' hierarchy, Wikipedia.

1 Introduction : Ontology Learning

Ontologies are an extremely essential approach mainly used in order to represent acquired knowledge. The ontology of a certain domain is about all essential concepts of it, their specifications, their hierarchies, whatever relations they have, and the axioms that constraint their behaviour [1]. The greatest challenge to use ontologies is the Semantic Web. It should be noted that the success of this new Web generation is above all dependent on the proliferation of ontologies, which require speed and simplicity in engineering them [2].

However, ontology engineering is a tough exercise which can involve a great deal of time and considerable costs. The need for (semi) automatic domain ontologies' extraction has thus been rapidly felt by the research world. Ontology learning is then the research realm referred to. As a matter of fact, this field is the automatic or semi-automatic support for the ontology engineering. It has indeed the potential to reduce the time as well as the cost of creating an ontology. For this reason, a plethora of ontology learning techniques have been adopted and various frameworks have been integrated with standard ontology engineering tools [3]. Since the fully automation of these techniques remains in the distant

future, the process of ontology learning is argued to be semi-automatic with an insistent need for human intervention.

Most of the knowledge available on the Web represents natural language texts [4]. Semantic Web establishment depends a lot on developing ontologies for this category of input knowledge. This is the reason why this paper focuses especially on ontology learning from texts. One of the still thorny issues of domain ontology learning is concepts' hierarchy building. In this paper, we are primarily involved in creating domain concepts' hierarchies from texts. We plan to use Wikipedia in order to foster the quality of our results. From this optics, literature reviews few research works dealing with this issue and none is making use of Wikipedia on the same way that it is harnessed in our approach.

In fact, Wikipedia is recently showing a new potential as a lexical semantic resource [5]. When this collaboratively constructed resource is used to compute semantic relatedness [6, 7] using its categories' system, this same system is also used to derive large scale taxonomies [8] or even to achieve knowledge acquisition [9]. The idea of harnessing Wikipedia plain text articles in order to acquire knowledge is quite promising. Our approach capitalizes on the well organized Wikipedia articles to retrieve the most useful information at all, namely the definition of a concept.

First, we will describe in Section 2 the ontology learning layer cake. In Section 3, we move straightforward to the explanation of our approach which will be followed by a corresponding evaluation in Section 4. Finally, Section 5 sheds the lights on some conclusions and research perspectives.

2 Ontology Learning Layer Cake

The process of extracting a domain ontology can be decomposed into a set of steps, summarized by [10] and commonly known as "ontology learning layer cake". The following page contains the figure which illustrates these steps.

The first step of the ontology learning process is to extract the terms that are of great importance to describe a domain. A term is a basic semantic unit which can be simple or complex. Next, synonyms among the previous set of terms should be extracted. This allows associate different words with the same concept whether in one language or in different languages. These two layers are called the lexical layers of the ontology learning cake. The third step is to determine which of the existing terms, those who are concepts. According to [10], a term can represent a concept if we can define: its intention (giving the definition, formal or otherwise, that encompasses all objects the concept describes), its extension (all the objects or instances of the given concept) and to report its lexical realizations (a set of synonyms in different languages).

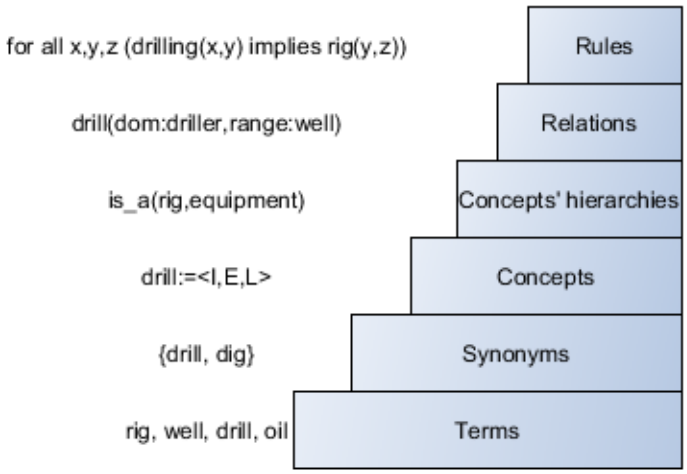

Fig. 1. Ontology learning layer cake (adapted from [10])

The extraction of concepts hierarchies, our key concern, is to find the relationship 'is-a', ie classes and subclasses or hyperonyms. This phase is followed by the non-taxonomic relations' extraction which consists on seeking for any relationship that does not fit in a previously described taxonomic framework. The extraction of axioms is the final level of the learning process and it is argued to be the most difficult one. To date, few projects have attacked the discovery of axioms and rules from text.

3 Concepts' Hierarchy Building Approach

Our approach tackles primarily the construction of concepts' hierarchies from text documents. We will make a terminology extraction using a dedicated tool for this task which is TermoStat [11]. The initial terms will be the subjects of a definitions' investigation within Wikipedia. Adapting the idea of the lexicosyntactic patterns defined by [12] to our case, the hyperonyms of our terms will be learned. This process is iterative which comes to its end when an in advance predefined maximum number of iterations is reached. Our algorithm generates in parallel a graph which unfortunately contains cycles and its nodes may have more then one hyperonym. The hierarchy we promise to build is the transformation result of the graph to a forest focusing on the hierarchic structure of a taxonomy. The figure on the following page gives the overall idea of the proposed approach.

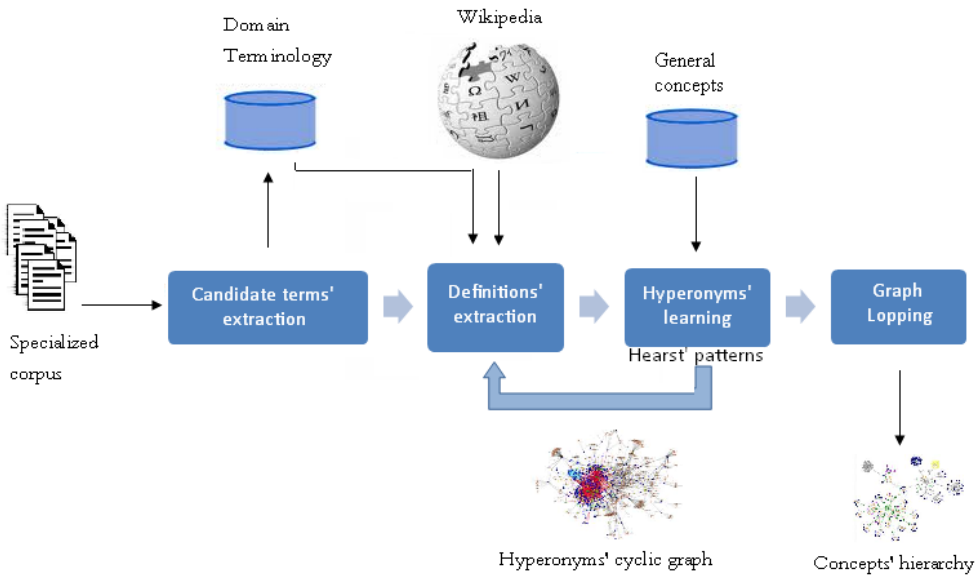

Fig. 2. Steps of the proposed approach

3.1 Preliminary Steps

In order to carry out our approach, we should first undergo the two lexical ontology learning's layers. The tool we used for the sake of retrieving the domain terminology is TermoStat. This web application was favored for determined reasons. In fact, TermoStat requires a corpus of textual data and, juxtaposing it to a generalized corpus such as BNC (British National Corpus), will give us a list of the domain terms that we need for the following step. Afterwards, we try to find out the synonyms among this list of candidate terms. The use of thesaurus.com as a tool in order to select synonyms was efficient. The third layer can be skipped in our context; concepts' hierarchies construction does not depend on the concepts' definitions. In other words, our algorithm needs mainly the candidate terms elected to be representative for the set of its synonyms (synset). The set of initial candidate terms is named $\mathcal{C}_O$.

3.2 Concepts' Hierarchy

The approach we are proposing belongs to two research paradigms, namely concepts' hierarchies construction for ontology learning and secondly the use of Wikipedia for knowledge extraction. The achievement of our solution relies heavily on concepts from graphs' theory.

a. Hyperonyms' Learning using Wikipedia

At the beginning of our algorithm, we have the following input data:

- $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ is an oriented graph such as $\mathcal{N}$ is the set of nodes and A is the set of arcs, $\mathcal{N} = \mathcal{C}_O$. Our objective is to extend the initial graph with new nodes and arcs; the former are the hyperonyms and the later are the subsumption links. The extension of $\mathcal{C}_i$, i is the iteration index, is done by using the concepts' definitions extracted from Wikipedia.
- $\mathcal{C}_{gen}$ is a set of general concepts for which we will not look for hyperonyms. These elements are defined by the domain experts including for example object, element, human being, etc.

S1 For each $c_j \in \mathcal{C}_i$, we check if $c_j \in \mathcal{C}_{gen}$. If it is the case, this concept will be skipped. Else, we look for its definition in Wikipedia. The definition of a given term is always the first sentence of the paragraph before the TOC of the corresponding article. Three cases may occur:

1. The term exists in Wikipedia and its article is accessible. Then we pass to the following step.
2. The concept is so ambiguous that our inquiry leads to the Wikipedia disambiguation page. In this situation, we ignore the word.
3. Finally, the word for which we seek a hyperonym does not exist in the database of Wikipedia. Here again, we skip the element.

S2 For the definition of the given concept, we apply the principle of Hearst's patterns. We attempt to collect exhaustive listing of the key expressions we need. For instance, the definition may contain: is a, refers to, is a form of, consists of, etc. This procedure permits us to retrieve the hyperonym of the concept c_j . The new set of concepts is the input data for the following iteration.

S3 Add into the graph $\mathcal{G}$ the nodes corresponding to the hyperonyms and the arcs that link these nodes.

b. From Graph to Forest

The main idea which shapes the following stage shares a lot with [13]. In fact, the graph which results from the preceding step has two imperfections. The first one is that many concepts are connected to more then one hyperonym. In addition, The structure of the resulting graph is patently cyclic which does not concord with the definition of a hierarchy. An adequate treatment is paramount in order to clean up the graph from circuits as well as multiple subsumption links. Thus, we will obtain, at the end, a forest respecting the structure of a hierarchy.

The following illustrative graph is a piece taken from the whole graph that we obtained during the evaluation of our approach. It represents a part of drilling wells' HSE namely the PPE (Personal Protective Equipment). The green rectangles are the initial candidate concepts.

The resolution of the first raised imperfection implies obviously the resolution of the second one. Therefore, we will use the following solution:

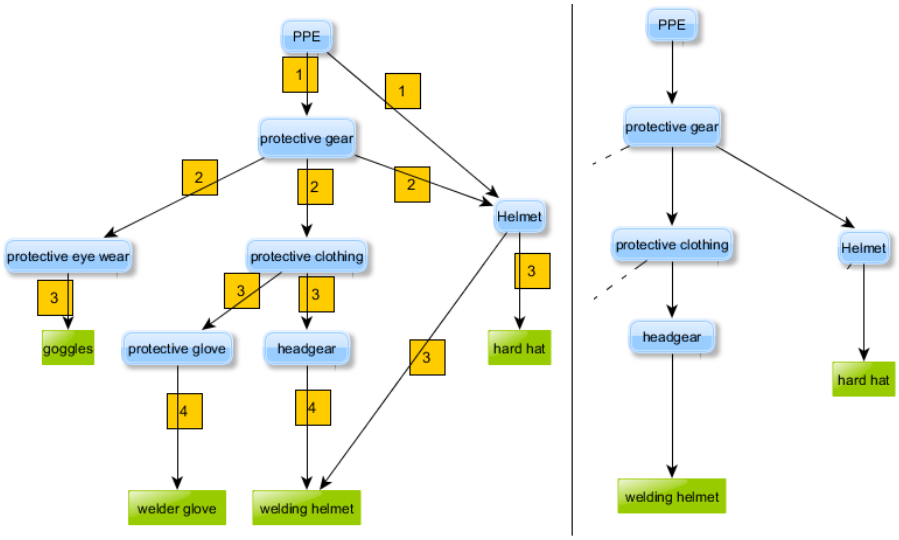

Fig. 3. From wells' drilling HSE graph to forest

1. Weigh the arcs as such as to foster long roads within the graph. We will increment the value assigned to the arc the more we go in depth (it is already done in fig.3).
2. We apply the Kruskal's algorithm[1956] which creates a maximal covering forest from a graph (fig.3).

Finally we have reached the aim we have planned.

4 Our Approach's Evaluation

Our evaluation corpus is a set of texts that are collected in the Algerian/British/Norwegian joint venture Sonatrach / British Petroleum / Statoil. This specialized corpus deals with the field of wells' drilling HSE . Throughout our approach, interventions from the experts are inevitable.
Tex2Tax is the prototype we have developed using Java. Jsoup is the API which allows us to access online Wikipedia. The same result is reached if using JWPL with the encyclopedia's dump. JUNG is the API we have used for the management of our graphs. The following page's figure is the GUI of our prototype.

The terminology extraction phase and the synonyms retrieving have given a collection of 259 domain concepts. The final graph is formed by 516 nodes and 893 arcs. After having done the cleaning, the concepts' forest holds 323 nodes, among them 211 are initial candidate terms. The amount of remaining arcs is of 322. In order to study the taxonomy structure we calculate the compression

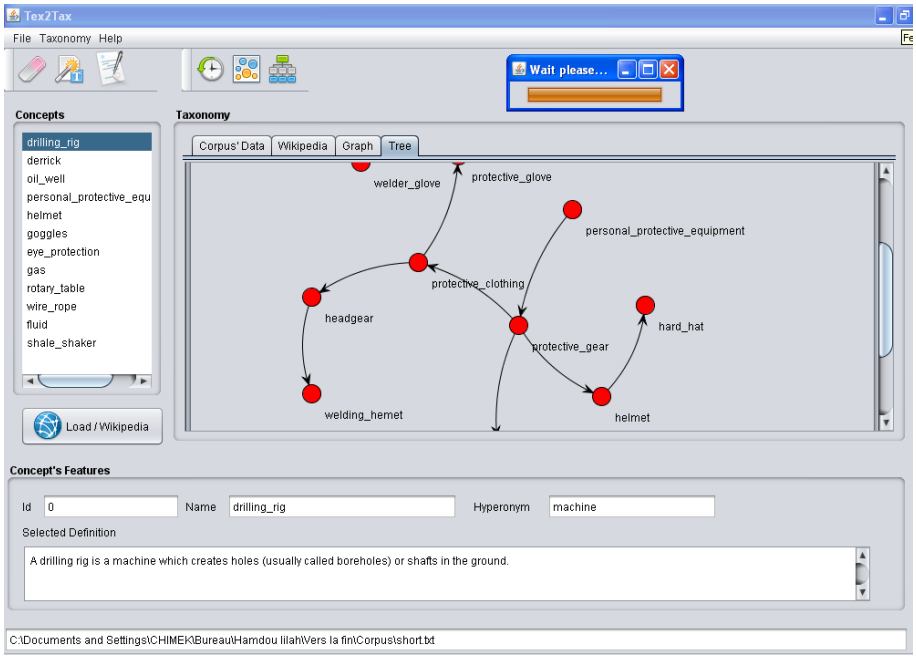

Fig. 4. Tex2Tax prototype's GUI

ratio for the nodes which is $0.63(323 = 516)$ and the one of the arcs which equals to $0.36(322 = 893)$.

$$LP = 0.63(323/516).$$
$$LR = 0.36(322/893).$$

The precision of our taxonomy is relatively low. This phenomenon is mainly due to the terms that do not exist in the database of Wikipedia. The graph's lopping is also responsible of some loss of nodes containing appropriate domain vocabulary.

5 Conclusion

Despite all the work which is done in the field of ontology learning, a lot of cooperation, many contributions and resources are needed to be able to really automate this process. Our approach is one of those few works that harness the collaboratively constructed resource namely Wikipedia. The results achieved and which are based on the exploitation of the idea of Hearst's lexico-syntactic patterns and the graphs' pruning is seen to be very promising. We intend to improve our work by addressing other issues such as enriching the research base

by the Web, exploiting the categories' system of Wikipedia in order to attack higher levels of the ontology leaning process such as non-taxonomic relations. Dealing with disambiguation pages of Wikipedia is of great value and multilingual ontology learning is, in addition, an alive research area which is just timidly evoked.

Acknowledgement We are thankful to the Sonatrach / British Petroleum / Statoil joint venture's President and its Business Support Manager for giving us the approval to access the wells' drilling HSE corpus.

References

[1] Cimiano,P., Mädche, A., Staab, S., and V ölker, J. Ontology Learning. In: S. Staab and R. Studer. Handbook on Ontologies. 2nd revised edition. Springer, 2009.

[2] IJCAI'2001 Workshop on Ontology Learning, Proceedings of the Second Workshop on Ontology Learning OL'2001, Seattle, USA, August 4, 2001. CEUR Workshop Proceedings, 2001.

[3] Mädche, A. Ontology Learning for the Semantic Web. Kluwer Academic Publishing, 2002.

[4] Zouaq, A. and Nkambou, R. A Survey of Domain Ontology Engineering: Methods and Tools, In Nkambou, Bourdeau and Mizoguchi (Eds): 'Advances in Intelligent Tutoring Systems', Springer, 2010.

[5] Zesch, Z., Müller, C., and Gurevych, I. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary . In Proceedings of the Conference on Language Resources and Evaluation (LREC). European Language Resources Association, 2008.

[6] Ponzetto,S.P.and M.Strube.Knowledge Derived from Wikipedia for Computing Semantic Relatedness. Journal of Artificial Intelligence Research 30, 2007.

[7] Strube M. et Paolo Ponzetto S. Wikirelate ! computing semantic relatedness using wikipedia. Proceedings of the National Conference on Arti cial Intelligence (AAAI), 2006.

[8]Ponzetto S. P. et StrubeM. Deriving a Large Scale Taxonomy from Wikipedia. AAAI '07, 2007.

[9] Nastase V. et Strube M.. Decoding Wikipedia Categories for Knowledge Acquisition. AAAI '08, 2008.

[10] Buitelaar, P., Cimiano, P., Magnini, B. Ontology learning from text: An overview. ontology learning from text: Methods, evaluation and applications. Frontiers in Artificial Intelligence and Applications Series 123, 2005.

[11] Drouin P., Acquisition automatique des termes : l'utilisation des pivots lexicaux specialises, thse de doctorat, Montral : Universit de Montral, 2002.

[12] Hearst M. A. et Schutze H. Customizing a lexicon to better suit a computational task. Proceedings of the ACL SIGLEX Workshop on Acquisition of Lexical Knowledge from Text, 1993.

[13] R. Navigli, P. Velardi, S. Faralli. A Graph-based Algorithm for Inducing Lexical Taxonomies from Scratch. Proc. of the 22nd International Joint Conference on Artificial Intelligence, 2011.