An Approach for Automatic Ontology Enrichment from Texts

Nassima Mellal, Tahar Guerram and Faiza Bouhalassa
Department of Mathematics and Computer Science, Larbi Ben Mhidi University, Oum El Bouaghi, Algeria
E-mail: nassima.mellal.univ@gmail.com, tahar.guerram@gmail.com, faiza.bouhalassa@gmail.com

Keywords: ontology, automaticEnrichment, NLP, SVO triplet, wordnet, texts, phytotherapy

Received: November 14, 2018

The automatic ontology enrichment consists of automatic knowledge extraction from texts related to a domain of discourse in the aim to enrich automatically an initial ontology of the same domain. However, the passage, from a plain text to an enriched ontology requires a number of steps. In this paper, we present a three steps ontology enrichment approach. In the first step, we apply natural language processing techniques to obtain tagged sentences. The second step allows us to reduce each extracted sentence to an SVO (Subject, Verb, and Object) sentence, supposed to preserve main information carried by the original sentence(s) from which it is extracted. Finally, in the third step, we proceed to enrich an initial ontology built manually by adding extracted terms in the generated SVO as new concepts or instances of concepts and new relations. To validate our approach, we have used "Phytotherapy" domain because of the availability of related texts on the WWW and also because its usefulness for pharmaceutical industry. The first results obtained, after experiments on a set of different texts, testify the performance of the proposed approach.

Povzetek: Predstavljena je metoda za izboljšave gradnje ontologij iz besedil.

1 Introduction

Ontology allows knowledge representation in graphical and intuitive manner but its construction and management is a hard task and a very time consuming operation. With the apparition of internet and new information and communication technologies, the mass of produced texts relating to different domains becomes huge and almost available for exploitation by interested users.

Hence, it would be very useful if this maintaining operation of ontologies will be done in an automatic or semi-automatic manner. This maintaining operation is sometimes called enrichment, sometimes it is called population as well as, but what is exactly the precise meaning of each one of this words? Ontology population is the process of inserting concept and relation instances into an existing ontology while ontology enrichment is the process of extending ontology, through the addition of new concepts, relations and rules [15]. As a main difference between the two processes is that ontology population preserve the ontology structure but ontology enrichment modifies it. Ontology learning is the process allowing the automatic generation of ontologies from a textual source called corpus. The ontology learning process is composed of several steps which are concept learning, taxonomic relation learning, non-taxonomic relation learning and finally axiom and rule learning.

We will interest, in the context of this paper, to the ontology enrichment process covering the three first steps of the ontology learning process, where we propose an approach for automatic ontology enrichment giving a text relating to a target domain. It is composed of three stages. In the first one, we use natural language processing techniques to extract sentences from text. Each extracted sentence is then annotated with part of speech tags and reduced to one or many binary relations (Subject, Verb, and Object) noted by SVO. The second stage consists of the determination of lexical relations (Hypernyms, hyponymy, synonymy,...) which may exist between the extracted terms (S, V and O) and the ontology concepts. For this purpose, we use an external knowledge source Wordnet. Finally in the third stage, the list of candidate's triplets (SVO) and lexical relations are used to enrich the initial ontology. To validate our work, we have chosen Phytotherapy as domain of discourse and the first results of precision, recall and f-measure metrics obtained are promising.

The remaining of this paper is organized as follows. Section 2 is devoted to the description of similar work, where we give recent and significant work in the field with their advantages and limitations. In Section 3, we

give detailed description of our ontology enrichment approach. Section 4 allows us to discuss the results obtained. Finally, we conclude our work and we give some perspectives in section 5.

2 Related work

Ontology is an explicit, formal specification of a shared conceptualization of a domain of interest [1]. New methods and tools are developed for reducing time and effort in the ontology construction process. The latter is called the ontology learning process. It is defined as the application of a set of methods and techniques in order to develop ontology from scratch or by enriching an existing ontology using different types of data: unstructured, semi-structured, and fully structured. In our context, we are interested in unstructured data, we speak about textual information.

Ontology learning from text is the process of identifying terms, concepts, relations, and optionally, axioms from textual information and using them to construct and maintain ontology. Techniques from established fields, such as information retrieval, data mining, and natural language processing, have been fundamental in the development of ontology learning systems. The ontology learning process is detailed in [15] (see figure 1)

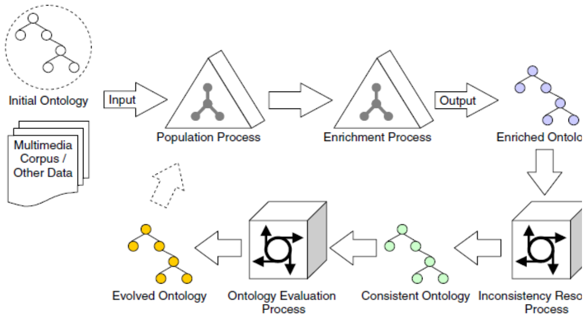


Figure 1: Ontology Learning Process [15].

According to ontology learning process, ontology enrichment is one of its important objectives. 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[5]. 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 can be divided into two main phases: the search for new concepts and relationships and the placement of these concepts and relationships within the ontology. According to [15], the process starts with the Concept Identification, then a taxonomy of concepts is constructed, the semantic relation extraction is the last step in enriching the initial ontology (see figure 2).

In [20], the process of enrichment is summarized within three main phases. The first is the **Extraction of representative terms in a specific domain.** It is the most important and difficult task. Several approaches (statistical and linguistic) are proposed for this aim. The second step concerns the **Identification of lexical relations between the terms.** Works in literature have focused on the identification of lexical relationships of hyperonymy, hyponymy, meronymy, synonymy and other more specific relationships that we call "transverse relations"[16],[22],[23]. The last phase aims to add the new terms as concepts/relations in relevant place in the ontology.

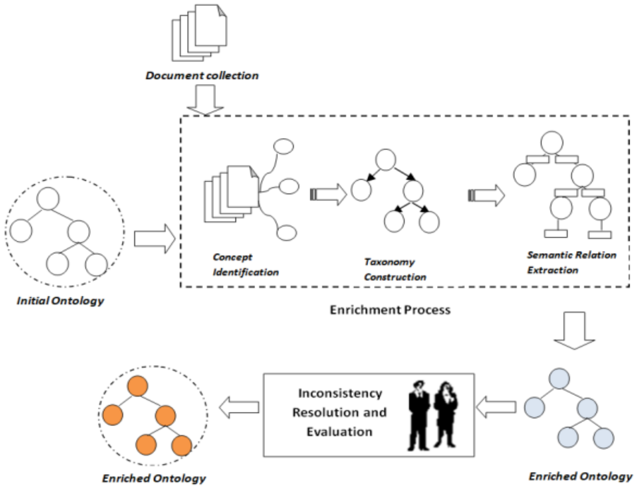


Figure 2: Ontology Enrichment Process [15].

In literature, different works of term extraction from textual corpus use two main approaches: **statistic analysis** and **linguistic analysis** approaches [17], [27], [28], [29]. The first one bases on statistic techniques of measures to facilitate the detection of new concepts and relations between them. Linguistic analysis uses linguistic techniques basing, generally, on detecting morphologic/ syntactic structures from the text in order to measure relativeness. Other works couple these two approaches and constitute an approach said « hybrid approach».

2.1 Statistical methods

They are often performed on large corpora. They are based on the co-occurrences, *TF*IDF*, *C/NC-value*, *T score*, *Dice Factor*, *Church Mutual Information*, and *word frequency* in the text in order to extract relevant terms to the target domain [2]. They base on the idea that is if two words coexist often in the same contexts, then they may be grouped together. This idea has been successfully realized in several works.

Drymonas employed C/NC-value to extract multi-word concepts [3]. Their proposed method “OntoGain”, aims to learn ontology from multi-word concept terms extracted from plain text. This method takes as input a corpus and produces a list of candidate multi-word terms, ordered by the likelihood of being valid terms, namely their C-Value measure [25]. NC-Value provides a method for the extraction of term context words (words that tend to appear with terms) and incorporates this information (from term context words) into the term extraction process [26]. OntoGain is applied on two separate data sources (a medical and computer corpus) the authors have evaluated 150 extracted terms with the help of domain experts. For computer science corpus, they obtained 86.67% precision and 89.6% recall, whereas for medical corpus, 89.7% precision and 91.4% recall were obtained.

Another example, of the statistical method, is the work of Mazari and his colleagues [4]. Their goal is to build ontology from a corpus of domain "Arabic linguistics". The process uses two statistical methods; the first is the “repeated segment”. It aims to identify the relevant terms that denote the concepts associated with the domain. The second is the “co-occurrence” method. It links these new extracted concepts to the ontology by hierarchical or non hierarchical relations. The first method performs an index of all words in the text by assigning a code corresponding to their positions in the corpus. Then it identifies all repeated segments in limiting itself to the same sentence. All of these segments are then filtered to remove unwanted segments and retain only those who are selected as candidate terms. The second method is based on the extraction of binary co-occurrents that meet one of the other more frequently than by chance and these two terms were included in the list found in the previous phase (detection of repeated segments). The co-occurrents will be selected with a frequency exceeding a statistically significant frequency due to chance. Then they will be compared with the labels of the ontology concepts. Terms may be added as new concepts, sub concepts or super concepts in the ontology and linked by Is-a or Part_of relation type. However, this approach is limited to Hyponymy and Meronymy relationships between concepts and the case, when both terms in the pair do not belong to the ontology labels, is not treated.

Therefore, these methods require human intervention for the positioning of the concepts in the ontology, or do not always identify the semantics of the relation, which influences on their accuracy.

2.2 Linguistic approaches

They use filtering techniques to manage text and to extract pieces of relevant information to the target domain. Works like those of Buitelaar and his colleagues [8] proposed a method mainly based on linguistics. It defines linguistic rules that extract concepts and relationships from collections of texts linguistically annotated. It is an approach that integrates linguistic analysis into ontological engineering. It supports the semi-automatic and interactive acquisition of ontologies from texts but also extension of existing ontologies. This methodology is associated with an OntoLT Protected plug-in [9] which uses predefined matching rules that automatically extract classes and candidate relationships from texts. For example, it maps the subject to a class, the predicate to a relation, the object's complement to a class, and creates the corresponding associative relationship between the two classes. If a rule is satisfied, the corresponding operators are enabled to create classes, relationships, or even instances that will later be validated and integrated into the ontology. The extracted ontology is integrated and can be explored in the Protégé development environment [9], which facilitates the management and sharing of the resulting ontologies. This approach has been used to build ontology in the field of neurology.

Other work in [6], aim at enriching an ontology from textual documents by relying on the linguistic analyzer "Insight Discoverer Extractor (IDE)". The analyzer outputs a tagged conceptual tree where each node carries a semantic tag attributed to the extracted textual unit based on the domain being processed. This approach presents a semi-automatic ontology population platform from textual documents. This platform provides an environment for matching linguistic extractions with the domain ontology of the client application using knowledge acquisition rules. These rules are applied, for each relevant linguistic label, to a concept, to one of its attributes or to a semantic relation between several concepts. They trigger the instantiation of these concepts, attributes, and relationships in the domain ontology knowledge base.

In [7], a linguistic method has been proposed in order to build domain ontology from Russian Text Resources. It uses a pipeline of linguistic methods (grafematic, morphological, syntactic and semantic analysis). *Grafematic analysis* is the initial analysis of the text on NL. It presents the input text data in a convenient format for further analysis (separation of input text into words, delimiters etc). *Morphological analysis* aims in construction of morphological interpretation of words of the input text (lemma, morphological part of speech…). *Syntactic analysis* is used for construction of syntactic tree from extracted syntactic groups consisting of sentences. *Semantic analysis* is used for building the semantic structure of one sentence. An algorithm of translating a syntactic tree into a semantic one applying a set of rules is proposed. As a result, the domain ontology can be built from the semantic trees extracted from text resources.

Linguistic approaches defining language rules (expressed as regular expressions) can identify specific terms associated with certain types of concepts in a domain.

The main limitations of rule-based approaches are that implementation requires a good knowledge of the field and requires manual work that is usually complex. In addition, rules are often defined for a specific domain or application and their application in other areas remains problematic.

2.3 Hybrid approaches

Combining linguistic information and statistical information is more commonly used to create term extraction modules. These hybrid systems use, first, linguistic filters to identify candidate terms, then statistical filters to distinguish terms from non-terms. In [10], an iterative method for semi-automatic acquisition of ontology and for enrichment of existing ontologies is proposed. It consists of a set of algorithms organized into modules aiming to extract concepts, relationships from texts. For the extraction of terms, a method based on statistical measures is applied to N-grams. A clustering method is then used to group these terms within concepts. The method proposes an algorithm for discovering Non-Taxonomic conceptual relations. It uses shallow text processing methods to identify linguistically related pairs of words, which are mapped to concepts using the domain lexicon. The algorithm analyzes statistical information about the linguistic output. Thereby, it uses the background knowledge from the taxonomy in order to propose relations at the appropriate level of abstraction. In this method, the conceptualization is automatic; it allows generating ontology automatically; the latter can then be refined and enriched with the help of an expert (adding new relevant concepts, removing irrelevant concepts).

A methodology implemented in the OntoLearn tool [13] provides different techniques for extracting ontological knowledge from texts. For the extraction of relevant terms from a domain, linguistic and statistical tools are combined to determine their distribution in the corpus. It also uses glossaries available on the Web. Lexical-syntactic patterns described by regular expressions are used to discover the subsumption relations between concepts. The internal structure of multiword terms is also used to extract this type of relationship, as in [8]. Using the WordNet lexical database also makes it possible to extract synonyms and other types of relationships.

In [11], another approach is developed to support the semi-automatic enrichment of ontologies from unstructured texts. It combines NLP and machine learning methods to extract new ontological elements, such as concepts and relations, from text. The method starts by identifying important parts of text and assigning them a set of basic ontological concepts from a given ontology. Then, it extracts new ontological concepts from these revealed pieces of text. Further, it determines hierarchical dependencies between these concepts by assigning them taxonomic relations. Finally, it creates ontological instances for the given ontology. These instances will be represented by concrete occurrences of some ontological concepts in a text document and will be linked by non-taxonomic relations. This method achieves F-measure up to 71% for concepts extraction and up to 68% for relations extraction.

In [14], automatic process for ontology population, from a corpus of texts, is proposed. It 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 first phase applies natural language processing techniques to identify instances of non-taxonomic relationships and properties of an ontology by annotating the inputted corpus. The second 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 aims to the classification of these instances, produces a populated ontology. Implementation of this process applied to the legal domain shows 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.

In [30] a process of ontology extension is proposed for a selected domain of interest which is defined by keywords and a glossary of relevant terms with descriptions. The methodology is semiautomatic, aggregating the elements of text mining and user-dialog approaches for ontology extension. Authors aimed to the analysis of business news by the means of semantic technologies. The methodology is used for inserting the new financial knowledge into Cyc [31], which maintains one of the most extensive common-sense knowledge bases worldwide.

In [33], a framework for enriching textual data is developed. It is based on natural language information extraction to include more structure and semantics. Authors implemented the proposed framework in a system, named Enrycher, which offers a user-friendly way to qualitatively enhance text from unstructured documents to semi-structured graphs with additional annotations. Since the system offers a full text enrichment stack, it makes the system simpler to use than having the user to implement and configure several processing steps that are usually required in knowledge extraction tasks.

According to the presented approaches, hybrid ones are the most adopted in the domain ontology learning process from texts. These different methods can be chained one by one to lead to better results [32]. But, the main drawback is that the majority of the methods, presented in this state of the art, do not take into consideration an important and preliminary step which can save time and resources. We speak about the automatic simplification / reduction of texts to be processed [20]. Developing a method, that led to reduce texts complexity and upgrades both readability and understandability by removing that which may be less important from texts, could improve and facilitate the enrichment ontology process.

3 Proposed approach

An important task of ontology learning is to enrich the vocabulary for domain ontologies using different sources of information. We propose an approach for automatic ontology enrichment giving a text relating to a target domain. First, a basic knowledge related to this target domain is predefined and represented in an initial ontology through a set of concepts and relationships between those concepts. The objective is to enrich this ontology by the content of texts relating to the same target domain through semantic analysis. As seen in the precedent section, generally, the essential steps in enrichment process are: *Extraction of terms, Identification of lexical relationships between terms* and *placing the extracted terms as Concepts/Relations in the existing ontology*(see figure 3)

3.1 Extraction of terms

One of our contributions, in this work, is the simplification of text in order to reduce its complexity. The majority of the proposed simplification methods rely on a set of manually defined transformation rules to be applied to sentences. In our approach, the proposed transformation rules are based on the segmentation of text into sentences and each sentence into tokens, each having its own POS (Part Of Speech). Then, based on these POS, we simplify, reduce and transform each sentence into a triplet SVO: (Subject, Verb, Object) supposed to carry the information of the sentences from which they are extracted. For this purpose, we use NLP techniques [18], [19]. The first step is divided into two sub-steps, we start first with parsing the text, and then we extract the significant terms.

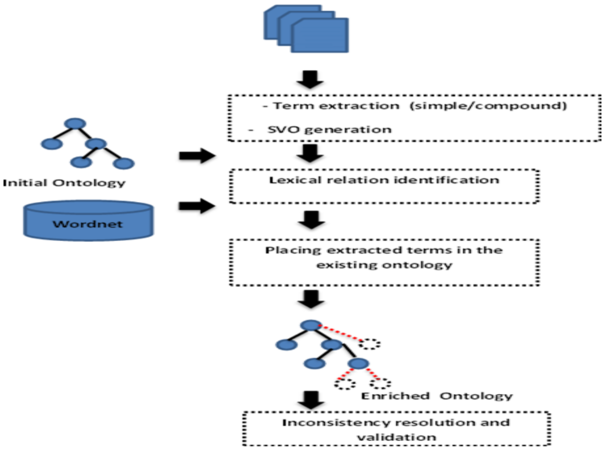

- Term extraction (simple/compound)
- SVO generation
Initial Ontology
Wordnet
Lexical relation identification
Placing extracted terms in the existing ontology
Enriched Ontology
Inconsistency resolution and validation

Figure 3: The proposed enrichment Process.

3.1.1 Syntactic analysis:

Syntactic Analysis or what we call preprocessing of texts. We aim, in this phase, to detect the type of words (verb, noun and adjective etc.), by *segmenting the text* into sentences. For each sentence, we extract its tokens having its own POS (Part Of Speech). These tokens may be simple or compound. In this last case, to make easy the detection of compound terms, we have proposed set of rules using English grammar [24] to define all possible compound terms (see the table bellow Table 1).

After term extraction, to simplify the sentences, stop-words will be removed from sentences. The stop words can be defined as words that don't have any remarkable importance. For example, *of, also, here, more, so, very,now.......*

The types of compound words	**Description**
NN + NNS	noun, common, singular or mass + noun, common, plural
NN +NN	noun, common, singular or mass + noun, common, singular or mass
NNS + NNS	noun, common, plural + noun, common, plural
NNP + NNP	noun, proper, singular+ noun, proper, singular
NN + NNP	noun, common, singular or mass + noun, proper, singular
JJ + NN	adjective or numeral, ordinal + noun, common, singular or mass
JJ + NNS	adjective or numeral, ordinal + noun, common, plural
NNP + NN	noun, proper, singular+ noun, common, singular or mass
NNP + NNS	noun, proper, singular+ noun, common, plural
NN + NN + NN	noun, common, singular or mass+ noun, common, singular or mass+ noun, common, singular or mass
NNPS + NNS	noun, proper, plural + noun, common, plural
NNPS + NN	noun, proper, plural +noun, common, singular or mass+ noun, common, singular or mass
VB + RB	verb, base form + Adverb
VB + RP	verb base form + Particle (up, off……etc)
MD + VB	modal auxiliary (could, should) + verb, base form

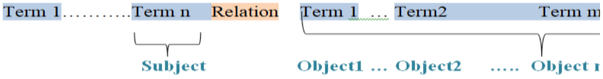
Table 1: Set of rules defining compound words.

3.1.2 Extraction and generation of SVO:

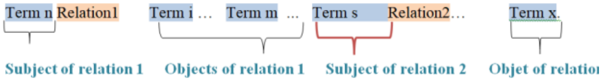
This step consists of simplifying, reducing and transforming each sentence of the text to a set of representative terms in the form of a triplet SVO: Subject, Verb, and Object. Here, we have to analyze all sentences obtained by text segmentation to a set of sentences. First, each sentence is annotated with POS tags and then the three parts of each sentence are delimitated: The subject part, the verbal part and the object part.

We have based, essentially, on the position of each term T (simple or compound) in the sentence S. To extract the relation in S, we test if the grammatical category of T is VB or VB + RB or VB + RP or MD + VB or VB+ADj (ADj: adjective situated directly after the verb), then T is the verbal part of the triplet. For example, in the sentence « The seed is rich in essential amino acidsand is used as cattle or poultry feed.” System detects two verbs *is_rich* and *use.* To extract subject and object parts, we distinguish the following cases:

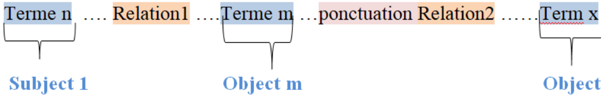
- If the sentence contains one verb, we select the nearest term before the verb as subject, and all terms after verb as objects.



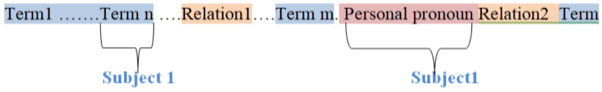
-
- in complex sentence containing more than one verb, for example according the next example, the subject of the second verb is term s



-
- For example, in « *Fresh Allspice berries when crushed can be mixed with a few drops of oil and massage onto the affected area to alleviate pain associated with rheumatism and arthritis.* » two subjects (*Allspice berries and pain)* for two verbs *(mixed* and *associated)*.
- If the sentence contains more than one verb and no term is before the verb 2, the subject of this latter is the same of the precedent verb.



-
- In the case where the personal pronoun is the subject of the sentence. We replace this pronoun by the subject of the precedent sentence.



-
- For example in "Asparagus is a climbing undershrub with widespread applications. It is useful in nervous disorders, dyspepsia, venereal diseases." The pronoun "it" is replaced by Asparagus.

Some kinds of sentences are not treated in the present work. For example, in the case of incomplete sentences those do not contain an object or subject. Also, in the case of negative sentences which are in negative form. At the end of this phase, we have a list of candidate's triplets (SVO) for enrichment.

3.2 Identification of lexical relationships between terms

In this step, we determine the relations which may exist between the extracted terms and the ontology concepts. For this purpose, we use an external knowledge source Wordnet [12]. Terms in WordNet are organized into synonym sets, called synsets, representing a concept by a set of words with similar meanings. Hypernyms, or the IS-A relation, is the main relation type in WordNet. Other types of relations are hyponymy, meronymy, synonymy, equivalence.

Several methods and applications focus on constructing taxonomic relationships rather than full-fledged formal ontologies. For that, our second contribution, in this work, is to develop an approach for the ontology enrichment taking in account taxonomic and non-taxonomic relationships between concepts. The achievement of this step depends on each candidate triplet SVO generated in the previous phase, and the set of concepts in the initial ontology. For each triplet and for each term T of this latter, we identify sets; each one is composed of words Wordnet having a lexical relation with the term T (hypernymy, hyponymy, synonymy). Subsequently, for each concept in the input ontology, we detect the lexical relation between this last and the term T. At the end, we have the types of lexical relations between the terms S, V, and O, and the concepts of ontology. How to place these terms in the ontology? This will be the subject of the next step.

3.3 Placing extracted terms in the initial ontology.

For each of the terms identified in step 1, we first check if it does not appear as a concept in the original ontology. In this case, our algorithm verifies possible approximations of meaning with the concepts of the ontology. The proposed enrichment process is illustrated in the algorithm below. It aims to add new concepts/relationships in the initial ontology. This must take into account the semantic links between concepts such as hyperonymy and hyponymy. The WordNet ontology is used for this purpose. For each triplet SVO, if the extracted term T exists in the initial ontology (IO), then no modification will be realized else, the following cases are distinguished:

Case 1: if the term T is a Subject or Object in the SVO, **then *if*** T is similar to an instance in the initial ontology (IO), so adding T as instance, ***else***, the following possibilities are distinguished (see figure 4),

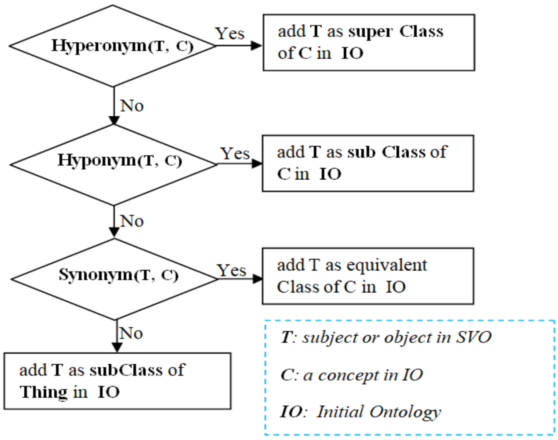

Hyperonym(T, C)
Yes
add T as super Class of C in IO
No
Hyponym(T, C)
Yes
add T as sub Class of C in IO
No
Synonym(T, C)
Yes
add T as equivalent Class of C in IO
No
add T as subClass of Thing in IO
T: subject or object in SVO
C: a concept in IO
IO: Initial Ontology

Figure 4: T is Subject or Object in IO.

Case 2: if T appears as a verb in the selected SVO, then as shown in figure 5, the following cases are distinguished.

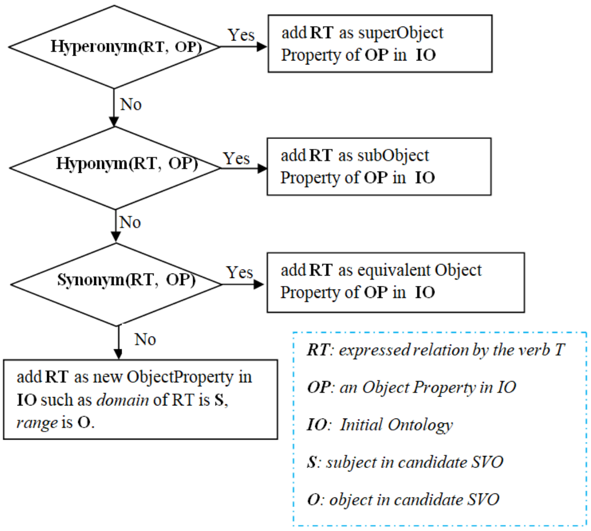

Hyperonym(RT, OP)
Yes
add RT as superObject Property of OP in IO
No
Hyponym(RT, OP)
Yes
add RT as subObject Property of OP in IO
No
Synonym(RT, OP)
Yes
add RT as equivalent Object Property of OP in IO
No
add RT as new ObjectProperty in IO such as domain of RT is S, range is O.
RT: expressed relation by the verb T
OP: an Object Property in IO
IO: Initial Ontology
S: subject in candidate SVO
O: object in candidate SVO

Figure 5: T is a verb in SVO.

Case 3: if the term T is a verb + Adjective, then the following alternatives are distinguished as shown in figure 6)

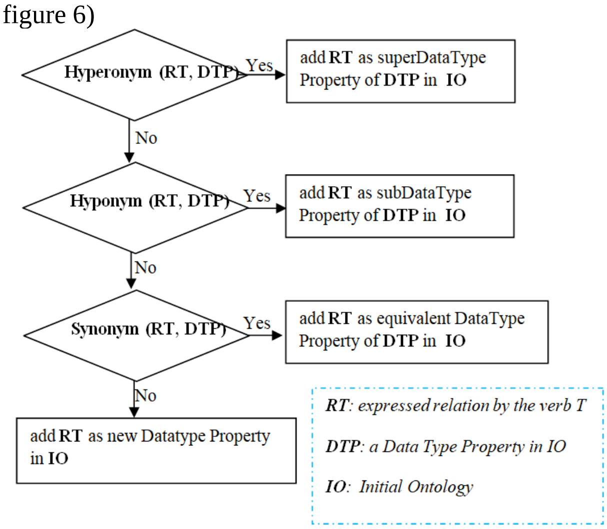

Hyperonym (RT, DTP)
Yes
add RT as superDataType Property of DTP in IO
No
Hyponym (RT, DTP)
Yes
add RT as subDataType Property of DTP in IO
No
Synonym (RT, DTP)
Yes
add RT as equivalent DataType Property of DTP in IO
No
add RT as new Datatype Property in IO
RT: expressed relation by the verb T
DTP: a Data Type Property in IO
IO: Initial Ontology

Figure 6: T is verb+adjective.

Indeed, enrichment here consists of adding concepts, instances, axioms and relationships. The following algorithm describes in detail the enrichment process.

```
Algorithm   Enrichment Process

Input: 1- SVO : list of triplet SVO
          2- IO: Initial ontology

BEGIN
Semantic_Relation ←Ø ;
For each  triplet in SVO DO
For each  term  T  in  triplet DO
LH←Hyperonyms list of T;
LHy←Hyponyms list of T;
LSy←Synonyms list of T;
For each  Entity E in  IO DO
IF  T   exists in  IO   THEN
      No modification
ELSE IF T (Subject or Object) AND (E is a  Class)
THEN   IF  (E  ∊ LH)  THEN
      Semantic_Relation← Hyperonym (E,T) ;
      IO←IO∪ (T  as a sub_ Class of E) ;
      ELSE  IF   (E  ∊LHy)    THEN
      Semantic__Relation← Hyponym (E,T) ;
      IO←IO∪ (T  as a Super_Class of E) ;
      ELSE  IF   (E  ∊LSy)    THEN
      Semantic__Relation← Synonym (E,T) ;
      IO←IO∪ (T  as  Equivalent_ Class of E ) ;
      ELSE
      IO←IO∪ (Concept  as Class) ;
      ENDIF
      ENDIF
      ENDIF
ENDIF
IF T(Subject or Object) AND(E is  Instance)
THEN IF   (E  ∊LSy)    THEN
      IO←IO∪ (T as  Instance) ;
      ENDIF
ENDIF
IF  T (Verb)  AND (E is  Object_ property)   THEN
      IF   (E  ∊ LH)   THEN
      IO←IO∪ (T as  Sub_Object_property) ;
      ELSE  IF   (E  ∊LHy)  THEN
      IO←IO∪ (T as  Super_Object_property) ;
      ELSE  IF   (E  ∊LSy)  THEN
      IO←IO∪( T as Equivalant_Object_property);
ELSE                    /* non-taxonomic relation*/
      IO←IO∪ (Relation  asObject_property) ;
      ENDIF
      ENDIF
      ENDIF
ENDIF
      IF   T (Verb + Adj.) AND  (E is
Data_Type_property)     THEN
      IF   (E  ∊ LH)   THEN
      IO ←IO∪(T as  Sub_Data_Type_Property) ;
      ELSE  IF     (E  ∊LHy)  THEN
      IO ←IO∪(T as  Super_Data_Type_Property) ;
      ELSE  IF    (E  ∊LSy)  THEN
      IO ←IO∪(T as
Equivalent_Data_Type_Property) ;
      ELSE
      /* non-taxonomic relation*/
      IO←IO∪(Relation as Data_Type_property) ;
ENDIF  ENDIF ENDIF ENDIF
ENDFOR   ENDFOR ENDFOR
END
OUTPUT :enriched ontology ;
```

4 Experiments and results

In this paper, we attempt to evaluate the performance of our proposed automatic ontology enrichment approach. We use **Phytotherapy** which consists of the use of plant derived medications in the treatment and prevention of disease. The World Health Organization (WHO) encourages the integration of the Phytotherapy in the health system [21]. However, the informal nature of its

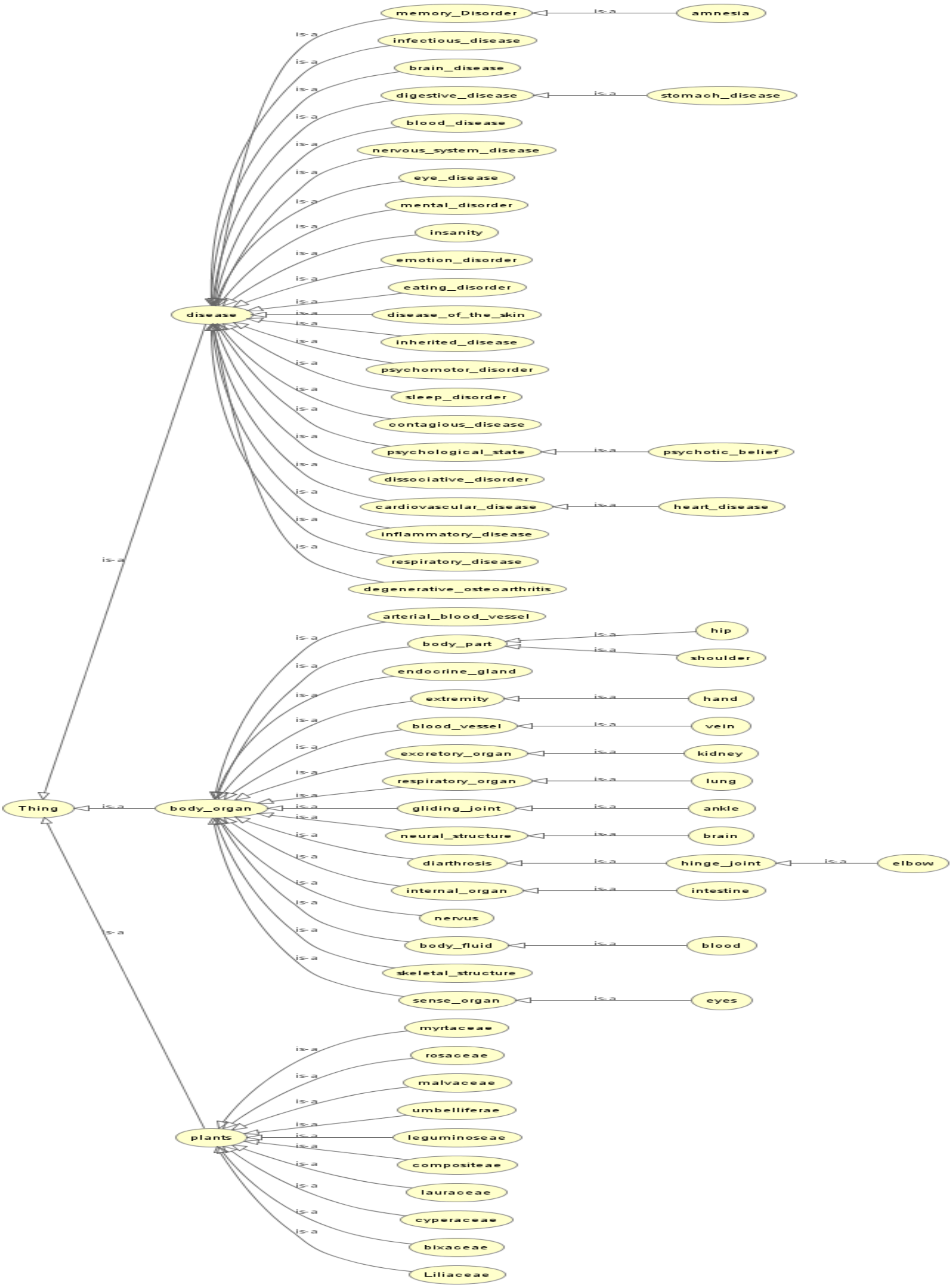


Figure 7: Initial Ontology of Phytotherapy Domain.

content makes difficult its use and practice. Our objective is not only formalizing the content of this medicine by means of ontology but also managing this latter automatically enriched from domain texts. This allows the end-users to be permanently informed about medicinal plants and their natural remedies against different diseases.

The initial ontology developed for the domain Phytotherapy describes some diseases; each disease belongs to a particular organ of the human body. It also lists the different plants that can cure diseases. For this purpose, we have defined three main classes in the OWL ontology: **Plant**, **Disease** and **Human_Organ.** (see figure 7).

To begin the enrichment process, we have used 25 texts, including 17 075 words speaking about three plants: Ginger, Aloe and Strophanthus. After applying the enrichment process, we obtain the following results (see table 2)

25 texts		
Expert SVO	2475	
Extracted SVO by system	1875 are true	375 are false

Table 2: Extracted SVO by Expert/System.

For example, in the segment of text : “Ginger also shows promise for fighting cancer, diabetes, non-alcoholic fatty liver disease, asthma, bacterial and fungal infections, and it is one of the best natural remedies available for motion sickness or nausea.” The generated SVO, are the following:

1. ginger_fighting_cancer
2. ginger_fighting_diabetes
3. ginger_fighting_non-alcoholic fatty liver
4. ginger_fighting_disease
5. ginger_fighting_asthma
6. ginger_fighting_fungal infections

The system takes these SVO one by one and enriches the initial ontology as following:

- *ginger* is added as an individual (instance) of the concept *Umbelliferae* (appearing as a class in the original ontology),
- *non-alcoholic fatty liver*, *fungal infections* are added as subclasses of *disease* Class,
- *asthma* is an existing individual (instance) of *respiratory disease*.
- The verb *fighting* is added as a relation between *Umbelliferae* concept and (*disease, fungal infections*, *asthma* and *non alcoholoic fatty liver*, *cancer* and *diabetes*) concepts, see the following figures (figure 8, figure 9 and figure 10).

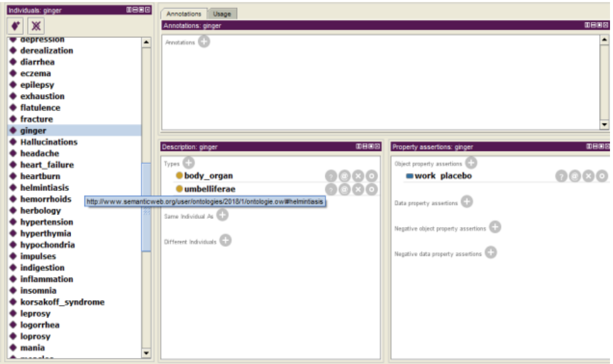


Figure 8: Creation of instance (ginger) from SVO to initial ontology (Protégé Window).

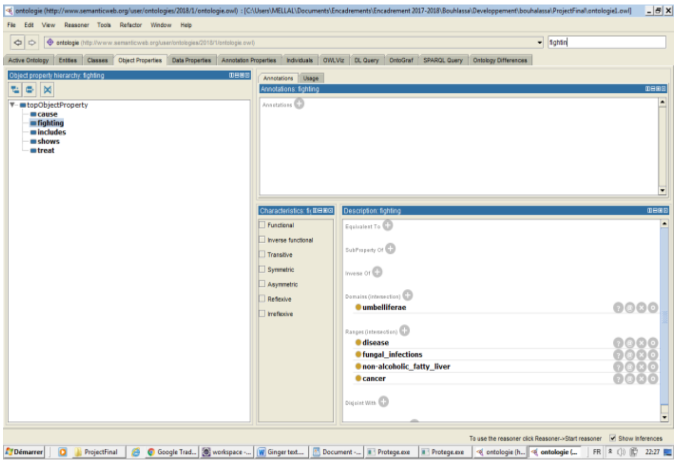


Figure 9: Placing Concepts/ relation from SVO to initial ontology (Protégé Window).

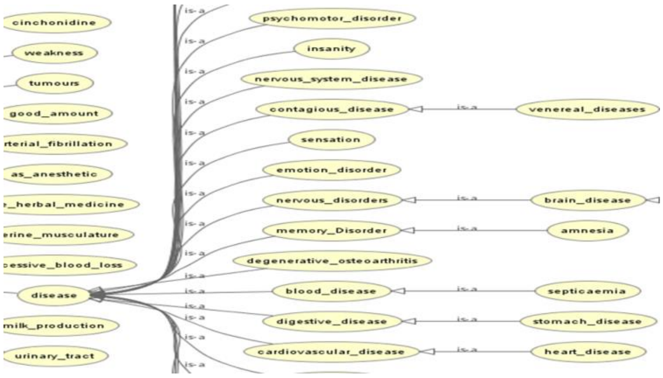


Figure 10: Part of enriched Ontology.

The consistency test and validation of the enriched ontology is done using Fact++ tool basing on the class/properties description. Here, the context of our work is only limited to first order relations.

To evaluate the performance of the proposed enrichment process, we use the precision, recall and F1 measure as follow:

Precision = (Number of generated SVO placed in the correct place in ontology by system/ Number of the generated SVO by system)

Recall = (Number of generated SVO placed in the correct place in ontology by system / Number of correct generated SVO by expert)

F-mesure = 2*Precision*Recall /Precision + Recall

Among the test set of 25 texts, extracted SVO by human experts agreed upon 2475 SVO. But after testing, the system gives 2250 SVO. In these 2250 SVO, 1875 SVO are correct and their terms are inserted in relevant places in the initial ontology. The implementation of the process shows results 83% as precision 75% as Recall and 78% as F-measure.

We have remarked that the proposed approach performs better with texts more than others. This is due to the type of sentences composing these texts. In fact, system gives best results in the case of verbal sentences containing a verb as a main part.

5 Conclusion and future work

In this paper, we have proposed an approach for automatic enrichment of a basic ontology composed of three stages. The first stage consists of applying natural language processing techniques to obtain tagged sentences. In the second stage, we reduce each sentence to a verbal one, called SVO (Subject, Verb, Object) sentence. Finally, in the third stage, we proceed to enrich an initial ontology built manually by adding new concepts, new relations and/or instances of concepts. We have distinguished three different approaches for automatic ontology enrichment: statistic based approach, natural language processing based approach and hybrid approach, which combines the two first approaches. The common problem of these approaches is that they don't reduce compound and complex sentences to their simplified forms before ontology enriching operation, which affect negatively their performance. Our approach is based on natural language processing techniques but augmented by a heuristic algorithm allows reducing extracted sentences to SVO (Subject, Verb, and Object) simple ones. This reducing step is very important because it allows improving the enrichment process performance. Another advantage of our approach is that it takes into account all types of relations, taxonomic and non-taxonomic, which allows us to have a good ontology enrichment rate.

To implement our approach, we have used a set of technologies proposed by the Semantic Web community (OWL, OWL-API, Wordnet ...) and the domain of natural language processing (Stanford Core NLP...). We have used Phytotherapy as domain of expertise since it is very important for pharmaceutical industry and as huge quantity of texts speaking about exits on the WWW. The first results obtained of precision, recall and f- measure are very encouraging (83% of precision, 75% of recall and almost 78 % of F-measure).

For this aim, some guidelines are to be taken into account. *First*, a survey of text segmentation and tagging algorithms must be done in the aim to use the most efficient ones. *Second*, treat the remaining cases of composed sentences and write the process of reducing texts in SVO in the form of an algorithm and try to optimize it. The *third* and the last guideline concerns the step of identifying SVO relationships with those of existing ontology and the placement of new concepts in it, this plays its preponderant role in the performance of the entire system, which is why a study and comparison of different ontology reasoners is imperative in order to use the most efficient one.

As future work, we plan first to enhance the performance of our approach by evaluating and improving the proposed algorithm. Also, we plan to extend our process using textual corpus to ensure that texts are in the domain we are interested in. Another future work consists of defining other new metrics like for example enrichment rate and enrichment efficiency metrics to measure the utility of our approach.

6 References

[1] T. R. Gruber . A Translation Approach to Portable Ontologies.Knowledge Acquisition, 5(2):199220, 1993
https://doi.org/10.1006/knac.1993.1008

[2] V.T. Nguyen. Méthode d'extraction d'informations géographiques à des fins d'enrichissement d'une ontologie de domaine. Doctoral Dissertation, Pau University (France), 2012.

[3] Drymonas,E., Zervanou,K. and Petrakis,E.G. Unsupervised ontology acquisition from plain texts: the Ontogain system. In: *NLDB*. Springer, Cardiff, United Kingdom. 2010.
https://doi.org/10.1007/978-3-642-13881-2_29

[4] A.C. Mazari , H. Aliane and Z. Alimazighi. Automatic construction of ontology from Arabic texts. ICWIT, pp. 193-202. 2012.

[5] N. Astrakhantsev, D. Fedorenko, D. Turdakov. Automatic Enrichment of Informal Ontology by Analyzing Domain-Specific Texts Collection". Materials of International Conference "Dialog", vol. 13, no. 20, pp. 29–42. 2014.

[6] F. Amardeilh, P. Laublet, J. L. Minel, "Document Annotation and Ontology Population from Lingusitic Extraction", Proceedings of Third International Conference on Knowledge Capture. 2005 .https://doi.org/10.1145/1088622.1088651

[7] Yarushkina, N.; Filippov, A.; Moshkin, V.; Egorov, Y. Building a Domain Ontology in the Process of Linguistic Analysis of Text Resources. *Preprints* 2018, 2018.
https://doi.org/10.20944/preprints201802.0001.v1

[8] P. Buitelaar, D.Olejnik, M. Sintek. A Protege Plug-in for Ontology Extraction from Text Based on Linguistic Analysis. In Proceedings of the 1st European SemanticWeb Symposium(ESWS). 2004.
https://doi.org/10.1007/978-3-540-25956-5_3

[9] H. Knublauch, R.Fergerson,.NF, Noy, M.A Musen. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. McIlraith, S.A., Plexousakis, D., Harmelen, F. van (Eds.), The Semantic Web – ISWC, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 229–243. 2004.
https://doi.org/10.1007/978-3-540-30475-3_17

[10] A. Maedche, S. Staab, N.Stojanovic, Y. Sure,R.Studer. Semantic portAL - The SEAL approach. In: Spinning the Semantic Web. MIT Press, pp. 317–359. 2001.
https://doi.org/10.1007/3-540-45754-2_1

[11] Ivana Lukšová. Ontology Enrichment Based on Unstructured Text Data. Master Thesis, Prague. 2013

[12] C. Fellbaum. WordNet: An Electronic Lexical Database. MIT Press, 1998.

[13] P. Velardi, P. Fabriani, M.Missikoff. Using Text Processing Techniques toAutomatically Enrich a Domain Ontology. In Proceedings of the InternationalConference on Formal Ontology in Information Systems, FOIS '01.ACM, New York,

NY, USA, pp. 270–284. 2001. https://doi.org/10.1145/505168.505194

[14] C. Faria, I. Serra, and R. Girardi, “A domain-independent process for automatic ontology population from text, Elsevier , Journal of Science of Computer Programming vol.95 pp 26–43. 2014. https://doi.org/10.1016/j.scico.2013.12.005

[15] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara and E. Zavitsanos, Ontology Population and Enrichment: State of the Art, Berlin/ Heidelberg, pp 134-166, Springer. 2011 https://doi.org/10.1007/978-3-642-20795-2_6

[16] Z. Sellami. Gestion dynamique d'ontologies à partir de textes par systèmes multi agents adaptatifs. Thesis Paul Sabatier University. 2012.

[17] A. Gomez Pérez, D. Manzano Macho. An overview of methods and tools for ontology learning from texts. The Knowledge Engineering Review, Vol. 19:3, 187–212, Cambridge University Press. 2005. https://doi.org/10.1017/S0269888905000251.

[18] D.Jurafsky, J.H.Martin. The Representation of Sentence Meaning. https://web.stanford.edu/~jurafsky/. 2018

[19] D. Jurafsky, J.H. Martin. Computing with Word Senses, https://web.stanford.edu/~jurafsky/. 2018

[20] M.Shardlow. A Survey of Automated Text Simplification. (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Natural Language Processing. 2014. https://doi.org/10.14569/specialissue.2014.040109

[21] M. Smith, A. Burton, T. Falkenberg. World Health Organization : Traditional Medicine Strategy 2014-2023”. 2014.

[22] N. Sheena, M.J Smitha, J. Shelbi. Automatic Extraction of Hypernym & Meronym Relations in English Sentences Using Dependency Parser. 6th International Conference On Advances In Computing & Communications. ICACC 2016, Cochin, India https://doi.org/10.1016/j.procs.2016.07.269

[23] M. Khodak, A. Risteski, C. Fellbaum, S. Arora. Automated WordNet Construction Using Word Embeddings. Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications. pages 12–23, Valencia, Spain. 2017 https://doi.org/10.18653/v1/w17-1902

[24] https://www.ef.com/ca/english-resources/english-grammar

[25] N. Hernandez. Ontologies de domaine pour la modélisation du contexte en recherche d’information. Thèse de Doctorat à l’Université Paul Sabatier France. 2006

[26] F. Rousselot et P. Frath. Terminologie et Intelligence Artificielle. 12èmes rencontres linguistiques, Presses Universitaires de Caen. 2002.

[27] A. Imsombut and J. Kajornrit. Comparing Statistical and Data Mining Techniques for Enrichment Ontology with Instances. Journal of Reviews on Global Economics, 6, 375-379. 2017 https://doi.org/10.6000/1929-7092.2017.06.39

[28] M.N. Asim, M. Wasim, M.U.G Khan, W. Mahmoud and H. Abbasi. A survey of ontology learning techniques and applications. Database, 1–24. 2018. https://doi.org/10.1093/database/bay101

[29] W. Wong. Ontology Learning from Text: A Look Back and into the Future. Article *in* ACM Computing Surveys. Volume 44 issue 4 pp 1- 36. 2012. https://doi.org/10.1145/2333112.2333115

[30] Novalija Inna, Mladenić Dunja. Ontology Extension Towards Analysis of Business News. Informatica Journal. Volume 34, N°4, pp 517—522. 2010.

[31] Cycorp, Inc., http://www.cyc.com

[32] Dunja Mladenić, Marko Grobelnik. Automatic Text Analysis by Artificial Intelligence. Informatica Journal, volume 37, N°1, pp 27—33. 2013.

[33] Tadej Štajner, Delia Rusu, Lorand Dali, Blaž Fortuna, Dunja Mladenić and Marko Grobelnik (2010), « A Service Oriented Framework for Natural Language Text Enrichment”, Informatica Journal, Volume 34, N°3, pp 307–313. 2010.