Wikipedia HTML Structure Analysis for Ontology Construction

Rim Zarrad*, Narjes Doggaz**, and Ezzedine Zagrouba***

*University of Manouba, Higher Institute of Documentation, Laboratory LIMTIC, Ariana, Tunisia, <zarrad_rim@yahoo.fr>

** University of Tunis El Manar, Faculty of Sciences of Tunisia, Laboratory LIPAH, Tunis, Tunisia, <narjes.doggaz@fst.rnu.tn>

*** University of Tunis El Manar, Higher Institute of Computer Science, Laboratory LIMTIC, Tunis, Tunisia, <ezzedine.zagrouba@gsn.rnu.tn>

Rim Zarrad is currently a university assistant at the *Higher Institute of Documentation of Tunis*. She received an engineering degree in computer science in 2005, a master's degree in 2007, and PhD in 2017 from the Faculty of Sciences of Tunis (FST). She is a member of the research laboratory in Computer Science, Modeling and Processing of Information and Knowledge (LIMTIC), Higher Institute of Computer Science of Tunis (ISI), University of Tunis El Manar (UTM). Her scientific research and production activities focus on ontology learning, data retrieval and knowledge extraction.

Narjes Doggaz received a BS degree in computer science from the High Institute of Management of Tunis, Tunisia, in 1986. She obtained her master's degree from the University of Nancy1, France, in 1987 and her PhD from the University of Nancy1, France, in 1992. She works on multi-agent systems and knowledge extraction. She is an assistant professor at Tunis El-Manar University and a member of LIPAH Research Lab in Tunisia. She is involved in different R&D projects and supervision activities.

Ezzeddine Zagrouba is Vice President of the Virtual University of Tunis (UVT). He is Fellow of an HDR in computer science, the founding director of the research laboratory in Computer Science, Modeling and Processing of Information and Knowledge (LIMTIC), Higher Institute of Computer Science of Tunis (ISI), University of Tunis ElManar (UTM), and Vice President of the Tunisian Research Association for Image Sciences (ArtsPi). His scientific research and production activities focus on information processing, artificial intelligence, computer vision and knowledge extraction. He has been the scientific coordinator of several projects and has participated in numerous national scientific and strategic commissions.

Zarrad, Rim, Narjes Doggaz, and Ezzedine Zagrouba. 2018. "Wikipedia HTML Structure Analysis for Ontology Construction." *Knowledge Organization* 45(1): 108-124. 43 references. DOI:10.5771/0943-7444-2018-2-108.

Abstract: Previously, the main problem of information extraction was to gather enough data. Today, the challenge is not to collect data but to interpret and represent them in order to deduce information. Ontologies are considered suitable solutions for organizing information. The classic methods for ontology construction from textual documents rely on natural language analysis and are generally based on statistical or linguistic approaches. However, these approaches do not consider the document structure which provides additional knowledge. In fact, the structural organization of documents also conveys meaning. In this context, new approaches focus on document structure analysis to extract knowledge. This paper describes a methodology for ontology construction from web data and especially from Wikipedia articles. It focuses mainly on document structure in order to extract the main concepts and their relations. The proposed methods extract not only taxonomic and non-taxonomic relations but also give the labels describing non-taxonomic relations. The extraction of non-taxonomic relations is established by analyzing the titles hierarchy in each document. A pattern matching is also applied in order to extract known semantic relations. We propose also to apply a refinement to the extracted relations in order to keep only those that are relevant. The refinement process is performed by applying the transitive property, checking the nature of the relations and analyzing taxonomic relations having inverted arguments. Experiments have been performed on French Wikipedia articles related to the medical field. Ontology evaluation is performed by comparing it to gold standards.

Received: 25 May 2017; Revised: 19 August 2017, 21 September 2017; Accepted: 6 December 2017

Keywords: taxonomic relations, concepts, extracted semantic relations, Wikipedia, ontology construction

1.0 Introduction

Knowledge organization systems are tools for classifying and managing knowledge and information within organizations. They cover all structured terminology systems such as thesauri, taxonomies and ontologies (Pieterse and Kourie 2014). Knowledge organization communities are particularly interested in the elaboration of ontologies (Almeida-Campos and Gomes 2017). A special feature of ontologies is that they include not only taxonomic relations among concepts but also known semantic relations (antonymy, meronymy, etc.) and some even more complex (labeled relations).

Ontology can be manually built by knowledge engineers and domain experts, resulting in tedious and cumbersome tasks. Moreover, manual acquisition of ontologies requires an extended knowledge of a domain and in most cases the result could be incomplete or inaccurate (Hazman et al. 2011). Since the automatic construction of ontologies constitutes a challenging task ontology construction has become an important research area. It refers to deriving ontological elements (high-level concepts and their relations) from input and building ontology from them (Shamsfard et al. 2003). Ontology construction aims at semi-automatically or automatically building ontologies from a given corpus with limited human intervention. A large collection of ontology construction methods from text has been developed in recent years. These methods focus on the extraction of an ontology's components, which are concepts and relations. The concepts are classes of objects grouped according to their properties, and relations are divided into taxonomic and non-taxonomic relationships. Taxonomic relationships are used to organize concepts hierarchically, whereas non-taxonomic relationships describe other kinds of interactions between domain concepts. Discovery of taxonomic and non-taxonomic relationships is the critical task in the ontology building process. This problem is largely studied in the literature and several tools have been developed. Classic methods for relationship discovery focus on text analysis and can be classified according to the chosen technique as distributional methods (Harris 1954; Shen et al. 2012) or syntactic methods (Morin 1999; Rios-Alvarado and Lopez-Arevalo 2013).

Distributional methods analyze the distributions of term co-occurrences in order to calculate a semantic distance between the concepts represented by those terms. Harris' hypothesis, which is the basis of word-space models, states that words that occur in similar contexts often share related meaning. These methods are robust and do not require preliminary knowledge of the field. However, as they disregard the context of sentences, these methods are not adapted for a precise analysis of the corpus. Moreover, these methods tend to extract conceptual links of different natures, which are difficult to dissociate without the contribution of an expert from the field. Syntactic approaches focus on the properties of the studied language. They set rules based on these properties in order to extract ontology components. However, it is not reasonable to specify a linguistic approach for each new study field. Most syntactic approaches focus on lexico-syntactic pattern-matching. They were introduced by A. Hearst in 1992 in order to extract synonym and hyponym relations from unrestricted text. Pattern-based approaches have been successfully used in the area of information extraction and ontology learning (Morin 1999; Panchenko et al. 2016; Barrière 2016). The lexico-syntactic patterns are rules describing a formed regular expression of words and grammatical categories corresponding to the syntactic forms of the relation and its arguments. These patterns characterize the semantics of the relation. Applying patterns to the free text gives a large number of extracted semantic relations. The main problem of this technique is the complexity and the diversity of the patterns which can express the same relation. Indeed, in the patterns-selection process, patterns must express properly the semantic relations to give better performance. The language used also has great impact on the ability of defining patterns and generalizing them.

Classic ontology construction methods often privilege analysis of the text. In the last decade, new approaches appeared (Sumida and Torisawa 2008; Paukkeri et al. 2012), which study the material structure of the documents to extract domain concepts. The structural organization of a document also conveys meaning and expresses hierarchical relations. These methods focus on titles, enumeration, definition sentence, underlined words, first and last part of a document, etc. In this context, we have proposed (Zarrad Doggaz and Zagrouba 2012; 2013) an approach to extract the main concepts and relations of the ontology by analyzing the structural meta-information of the document and especially the titles of each document. The concepts are filtered by analyzing the document structure (title, style and hyperlink). Taxonomic relations discovery is performed using both syntactic analysis and lexico-syntactic patterns-matching methods whereas non-taxonomic relations discovery is established by analyzing the titles hierarchy of each document. To retrieve non-taxonomic relations, we combine three methods: title pattern detection, three-level analysis and two-level analysis. The proposed methods extract not only taxonomic and non-taxonomic relations but also give the labels corresponding to each non-taxonomic relation.

In this paper, we describe an extended version of this approach. We propose to improve the two-level analysis and apply a refinement of the obtained relations set in order to keep only those that are relevant. The two-level analysis extracts non-taxonomic relations by considering titles belonging to two nested levels. Moreover, we propose to

apply pattern matching technique to enrich the set of non-taxonomic relations by searching such known semantic relations as synonymy, meronymy, antonymy and causality relations. Once the relations are extracted, a refinement step is performed in order to eliminate redundancy and remove invalid relations keeping, therefore, only those that are relevant. The refinement process is performed by applying the transitive property, checking the nature of the relations and analyzing taxonomic relations having inverted arguments.

The proposed approach generates an ontology composed by a set of concepts and semantic relations. The extracted relations can be taxonomic or labelled. We note that the built structure is semantically richer than a thesaurus. In a thesaurus, the relations between terms are either synonymy or taxonomy or association. We do not find in a thesaurus labeled relations such as "tuberculosis, rifadine, treatment." This kind of relation is found only in ontologies. To test and evaluate the proposed approach, we use a French corpus collected from Wikipedia articles related to the medical field. The evaluation of the extracted relations is performed using two repositories: the thesaurus MeSH and the category system of Wikipedia.

In the next section, we present a review of research dealing with relation extraction. The architecture of the proposed ontology construction process is presented in section 3. The methods for discovering non-taxonomic relations by analyzing titles are detailed in section 4. Section 5 concerns the refinement step. The experimental and comparative study is given in section 6. Finally, we give some conclusions, and we present some perspectives.

2.0 Related work

One of the fundamental tasks in ontology construction is the extraction of taxonomic and non-taxonomic relations between domain concepts. A taxonomic relation is also called a hyponym/hypernym relation. A noun phrase X is a hyponym of a noun phrase Y if X is a subtype or instance of Y (Snow et al. 2005). For example, from the concept "school," which is a subtype of "educational institution," a taxonomic relation can be established between "school" and "educational institution." Thus, "school" is a hyponym of "educational institution" and conversely "educational institution" is a hypernym of "school." A non-taxonomic relation represents an interaction between concepts. In other words, the two concepts A and B are linked by a non-taxonomic relation if A is semantically related to B. These relations can be active/passive relations, causal relations, locative relations, etc. A label is generally assigned to a non-taxonomic relation. Its role is to describe the relation between the two concepts. The less explicit and more complex use of words for specifying relations other than hypernymy causes the tasks of discovering and labeling non-taxonomic relations to be more challenging (Wong et al. 2012).

Methods dealing with relations discovery focus generally on the analysis of the text (Kermanidis and Fakotakis 2007). In the last decade, several approaches tried to build ontologies by analyzing the material aspects of the texts. Indeed, the visual properties of texts are not just an ornament of the text but constitute an important component implied in their significance. Several methods dealing with ontology construction focus on items belonging to the document structure. They analyze titles, page layout, character policy, etc. Among methods dealing with document structure, several focus on titles belonging to each document in order to extract ontology components. Generally, titles have an important role in the text. They are textual objects which are visible by typographical properties (bold, numbered, etc.) or dispositional ones (tabulation, line feeds before or after enumerations, etc.). These properties enable the titles to play an important role in the level of the material organization of the text. According to Jacques and Rebeyrolle (2006, 3), "The nested titles of sub-sections belonging to a given section reflect the nested relations existing between these sections." The text can be then considered not like a linear succession of paragraphs but like a structure of elements of high level, which include other elements. The document's structure can be viewed like a division into sections and sub-sections, usually equipped with a title. In addition to providing the segmentation and visual organization of the text, the titles contribute to the construction of semantic content.

We find in literature several methods that use Wikipedia articles as a corpus for relation extraction (Hazman et al. 2011; Medelyan et al. 2013; Filipiak and Lawrynowicz 2014). Indeed, Wikipedia is the largest and most popular general reference work. It is a semi-structured knowledge source that is rich in terms of semantic and lexical information. Moreover, Wikipedia articles are highly structured with titles that conform to specific guidelines (length, precision, etc.). This encyclopaedia contains not only structured texts but also explicit information such as infoboxes and categories. Most recent work focuses on exploiting categories and infoboxes from Wikipedia to establish relationships between domain concepts. Categories are a system of thematic classification of Wikipedia articles. They are found at the bottom of each article page. The categories were used in many ontology construction approaches (Ponzetto and Strube 2007; Suchanek et al. 2008). However, we note that category hierarchy in Wikipedia contains duplication and sometimes it is inconsistent compared to other manually created hierarchies like WordNet. Another widely used technique for extracting relations is the use of infobox (Wu et Weld 2007; Abd El-atey et al. 2012). An infobox is a preformatted table

of dynamic data that summarizes important information about a topic and that is located generally on the right of the article. The problem related to the use of this table is that the extracted knowledge is mostly limited to named entities with proper names, such as cities, persons, species, movies, organizations, etc.

Paukkeri et al. (2012) propose a method based on document structure in order to extract taxonomy. In this study, each Wikipedia article is specific to a domain concept. Three different approaches are proposed to reduce the dimensionality of the processed data. The first approach uses a combination of heuristic criteria exploiting the document structure (titles, underlined words, first and last part of a document) by means of fuzzy logic. The second is a language independent approach based on statistical analysis and key phrase extraction. The third approach is based on the classic measure TF-IDF (Joachims 1997). The Term Frequency TF(w,d) is the number of times the word "w" occurs in document "d" whereas the IDF factor indicates the opposite frequency of a document. The IDF factor aims at giving a more important weight to the least frequent terms considered as discriminant. We note that, in the approach proposed by Paukkeri et al., the document structure was used only for concept extraction and is not involved in relation extraction tasks.

Sumida and Torisawa (2008) propose to extract taxonomic relationships using hierarchical layouts in Wikipedia. These layouts concern title hierarchy, numbered lists and bulleted lists. This method also uses a machine learning technique, pattern-matching method and other existing methods for extracting relations from definitions and category pages. Their system extracts more than 1.4×10^6 taxonomic relationships from the Japanese version of Wikipedia with 75.3% as precision value.

Patrick Arnold (2014) also uses Wikipedia articles to extract semantic relations. However, he focuses more on the text of Wikipedia articles, especially Wikipedia definitions, rather than on the existing category system or infobox. His method covers not only taxonomic relations but also semantic relations such as "equal," "has-a" and "part-of." Precision values range between 74% and 81% with an average of 79%. Kazama and Torisawa (2007) consider the first sentence of a Wikipedia article as the definition sentence and extract the concepts from the definition sentence. They exploited syntactic patterns to identify the hypernym in the definition sentence.

Several approaches aim to extract relations between domain concepts by combining a document structure analysis with a syntactic method. Most of these approaches use pattern matching techniques. The work by Aussenac-Gilles and Kamel (2009) concerns ontology construction from XML database specifications. The approach initially consists of studying XML tags and their hierarchy to extract taxonomic and non-taxonomic relations between domain concepts. Moreover, the remainder of the text is analyzed using lexico-syntactic patterns. The corpus utilized contains very short and synthetic texts so the proposed approach is poor in terms of number of concepts and relations. Moreover, the same concept may be classified in different levels of the hierarchy. In other words, we find in the ontology different occurrences of the same concept. Each occurrence has its own properties. Hence, some extracted concepts and relations in this approach are inconsistent. In Snow et al. (2005), the approach proposed learns syntactic patterns for automatic hypernym discovery. Taxonomic relations are extracted automatically from text using dependency paths and WordNet. Barbu and Poesio (2009) propose the use of patterns to build taxonomies of concepts from raw Wikipedia text. The construction process starts with the generation of concept hierarchies from WordNet. The concepts in each hierarchy are then mapped onto Wikipedia pages. The authors assume that the concepts that are classified under the same node in a taxonomy are described in a comparable way in Wikipedia. Finally, lexico-syntactic patterns describing semantic structures are automatically learned.

Compared to extended work on relation extraction, little attention has been given to labelling non-taxonomic relations. Label extraction is the more intricate task in the ontology construction process. To our knowledge, works dealing with label extraction are usually based on syntactic analysis and particularly on the occurrences of verbs in sentences (Sanchez and Moreno 2008; Weichselbraun et al. 2009; Punuru and Chen 2012). In this paper, we present an approach that gives taxonomic and non-taxonomic relationships. For most non-taxonomic relationships, the proposed method also provides the corresponding labels. The originality of this approach is that it analyzes the document structure and primarily the title hierarchy in order to extract taxonomic or labelled relations.

3.0 Proposed approach for ontology construction

The proposed approach analyzes the text structure and especially the titles of the documents to extract the ontology concepts and relations. Our approach consists of four principle steps as described in Figure 1. The first step consists of the selection and pre-processing of the corpus documents by applying NLP tools. We focus then on the extraction of concepts of the ontology that correspond to the relevant terms of the field. The relations linking these concepts are extracted in the third step. Finally, we apply a refinement to these relations in order to keep only those that are relevant. The ontology construction steps are described in more detail in the following sections.

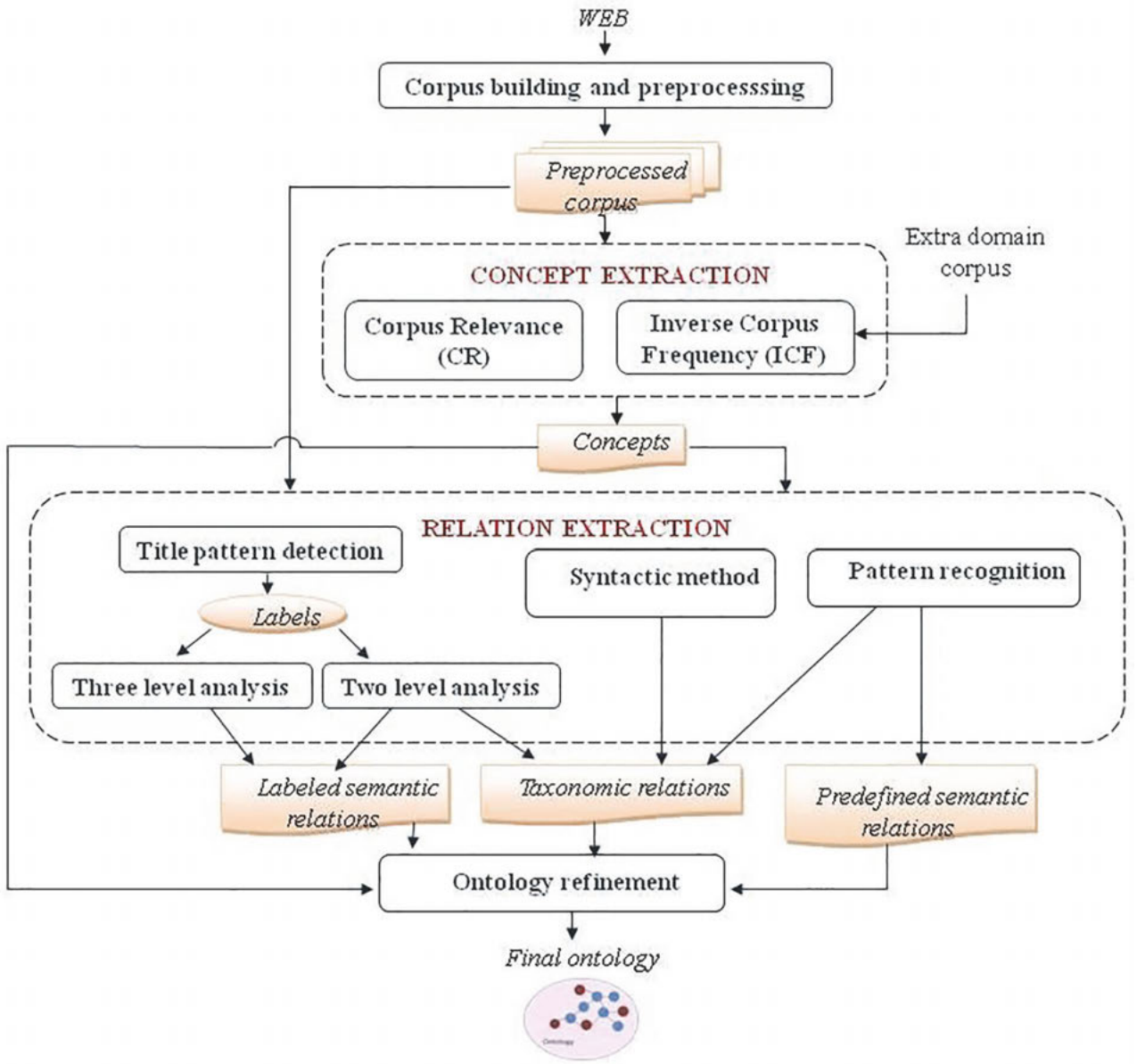

Figure 1. Ontolobgy learning process.

3.1 Corpus building and pre-processing

The first step of the proposed approach is the selection of the corpus documents. In the proposed approach, we use HTML Wikipedia documents. The selection of these documents is performed using Google Web Search API. It is a library tool given by Google programmers to extract the results of a request. A set of keywords describing the field followed by a set of search parameters is introduced. As a result, we obtain a set of URLs corresponding to the websites that are representative of the field and that verify the defined search criteria.

Once the corpus is built, we focus on the analysis of the collected documents using several natural language processing (NLP) tools. Text pre-processing is an essential part of our system, since the characters, words and sentences identified at this stage are the fundamental units passed to all further processing steps. It is composed of three steps: tagging and lemmatization, HTML parsing and stop-sets generation.

Tagging identifies the grammatical category of each word whereas lemmatization gives the canonical form of an inflected word. To perform these operations, we use the TreeTagger tool developed by Helmut Schmid (1994). The HTML parsing extracts data by analyzing the tags of the HTML documents. In our approach, we use HtmlParser 1.6, which is a free java library for the extraction and the text processing of a web corpus. The last step of corpus

pre-processing consists of the removal of the stop-words and the stop-titles that appear in the corpus. Indeed, textual documents contain generally many words that occur frequently in the texts but have no informational content (e.g., articles, prepositions, conjunctions). To remove these words, we define a general stop-set containing common and non-informative words. We also define a stop-title set to eliminate words that occur frequently in titles of general structured texts and especially in Wikipedia articles (references, external link, bibliography, etc.).

3.2 Concept extraction

After the preprocessing step, we focus on the extraction of terms that are candidates to be the concepts of the ontology. These terms are called candidate terms and are denoted by CT. They are linguistic units that qualify objects or notions of the real-world. In the proposed approach, we use a syntactic method that extracts not only simple terms but also nominal groups called nominal noun phrases. We are interested in the most abundant noun phrases containing specialized terminology. Thus, according to the canonical syntactic structure of the CT, we define two classes of noun phrases:

- One-Noun-phrase-Class (Cl_1): contains noun phrase composed of one word. This word is either a "Name" or a "Named_Entity;"
- Two-Noun-phrase-Class (Cl_2): composed by noun phrases containing two words and having as syntactic structure the sequence "Name + Adjective" or "Named_Entity + Adjective."

We notice that, in the French language, the adjective is usually located after the noun ("le chat noir") while for English, it is exactly the other way around ("the black cat").

Once the CT is extracted, a filtering process is applied in order to keep only the most important terms that correspond to the main concepts of the studied field. The concept extractor step is established by analyzing the structure of the text including titles, styles used and hyperlinks that appear in these documents. To measure the relevance of a CT in the corpus, we define the CR-ICF measure inspired from the classic TF-IDF measure (Joachims 1997). The proposed measure is the product of two factors CR (corpus relevance) and ICF (inverse corpus frequency).

Let C = {d_1, d_2, … , d_m} be a corpus of "m" HTML documents (d_j). The corpus relevance CR of each candidate term "S" is the sum of its normalized relevance (NR) in all the documents of the corpus. It is computed as follows:

$$CR(S) = \sum_{j=1}^{m} NR(S, d_j) = \sum_{j=1}^{m} \left(\frac{R(S,d_j)}{\sum_{k=1}^{n_S(j)} R(S_k,d_j)}\right) \quad \text{(eq. 1)}$$

where $n_s(j)$ is the number of CT in the document "d_j" and R(S,dj) is the relevance of the CT "S" in the document "dj." The denominator is the sum of the relevance of all the syntagms in this document.

The relevance of a CT "S" in a document "d_j" is defined as the sum of its "title" relevance $R_{Title}(S)$ and its "style" relevance $R_{Style}(S)$ in the document. In fact, titles, hyperlinks and words having bold or italic style generally represent elements carrying relevant information. We have focused on these elements in order to filter the candidate terms.

The normalized relevance of a CT "S" in the document "d_j" is used to avoid the problems related to the length of the document.

The ICF factor indicates the opposite frequency of the studied "corpus." It is based on the occurrences of the CT in other corpora of different fields in order to check whether the extracted CT is a general term or is specific to the studied field. A candidate term occurring frequently in other corpora will be considered as a general word (i.e., the term is not specific to the studied field) and consequently, it will be removed. The ICF measure aims at giving a more important weight to the least frequent terms in the other corpora, considering them as relevant. The ICF factor of a CT "S" is computed as follows:

$$ICF(S) = \log\left(\frac{|C|}{|C_j/S \in C_j|}\right) \quad \text{(eq. 2)}$$

where "C_j" is the "j^{th}" corpus and |C| is the total number of the corpora.

Once CR and ICF measures are computed for each CT, we calculate CR-ICF values. A filtering step is then applied and the CTs having a CR-ICF value higher than a given threshold are considered relevant and selected as concepts of the study field.

4.0 Relations extraction

The main advantage of the proposed approach is its ability to extract, in addition to taxonomic relations, those that are non-taxonomic. In the ontology, taxonomic relations play a role in classifying concepts from the most general to the most specific. In order to extract such types of relations, we propose two methods: a syntactic analysis method and a pattern matching technique. The first method focuses on the two classes of domain concepts, "Cl_1" and "Cl_2," in order to extract hierarchical relations. The concepts of the class "Cl_1" are composed of one word while those belonging to "Cl_2" are composed of two words. In the grammatical context, an adjective is a word

which modifies the meaning of the word that precedes it by making it more specific. Thus, from a concept C_1=N_1 belonging to "Cl_1" and a concept C_2=N_1+adj belonging to "Cl_2," a taxonomic relation is then defined between "C_1" and "C_2." As an illustrative example, let us consider the case of the medical concepts C_1="artery," C_2="pulmonary artery" and C_3="femoral artery." In anatomy, "pulmonary artery" and "femoral artery" are two types of the "artery" concept. Thus, two hierarchical relations are established between C_2="pulmonary artery" and C_1="artery" on the one hand and between C_3="femoral artery" and C_1="artery" on the other hand.

We propose also a pattern matching technique in order to extract taxonomic and non-taxonomic relations. For taxonomic relations extraction, our method searches for contexts that mark the hyperonymy relations. For this task, we use first the Marshman's patterns (Marshman 2008), which are specific to hyperonymy relation in the French language. Then, the relations are filtered in order to eliminate those that are invalid.

The pattern matching technique is also used to retrieve non-taxonomic relations. The synonymy, meronymy, antonymy and causality relations are the most frequently used by pattern matching works. We use the patterns defined respectively in Cartier (2015) and Morlane-Hondere and Fabre (2010) to extract synonym and antonym terms. The study of meronymy relation is performed by combining the patterns defined in Loffler-Laurian (1994) and Marshman et al. (2002). The patterns presented in Marshman et al. are also used to learn causal relations. The patterns used when applying this method usually give relevant results. Nevertheless, they are language specific and cannot be generalized to corpora belonging to other languages. Moreover, they deal with specific relations (synonymy, meronymy, antonymy, causality) and cannot easily generate domain relations. But, in the case of a domain corpus, we find several relations related to the studied field. For example, in the medical field, several terms are associated with the labels "symptom," "treatment," "side effects," etc. To learn such kinds of relations, we must define specific patterns (is a symptom of, is a treatment of, etc.). In other words, for each semantic relation, we should define the associated patterns. We conclude that this method depends on the studied domain and cannot be generalized.

To resolve this problem, we propose new methods that extract labeled relations and that are not dependent on the studied domain. The set of extracted relations is larger and semantically richer than that given by pattern matching techniques. We propose three methods based on title hierarchy that are pattern detection, three-level analysis and two-level analysis. These methods analyze the document structure, especially the title hierarchy of the documents. Indeed, the titles in a document are often chosen to express some organization of the text content. The hierarchy of titles corresponding to each document is then represented by a tree structure. The proposed approach has the advantage of extracting non-taxonomic relations as well as the associated labels.

We give below some notations which are used in the rest of this paper:

- R denotes the set of all extracted relations.
- TR corresponds to the set of "taxonomic relations" and NTR corresponds to the set of "non-taxonomic relations."
- A taxonomic relation between two concepts "$concept_1$" and "$concept_2$" is represented by the pair ("$concept_1$, $concept_2$"). These two concepts correspond to the arguments of the relation where "$concept_2$" is a sub-concept of the "$concept_1$."
- A non-taxonomic relation is represented by the triplet ("$concept_1$," "$concept_2$," "lab") where "lab" is the label of the relation linking the two concepts "$concept_1$" and "$concept_2$."
- TermSet(T) denotes the set of terms belonging to a title *T*:
$$TermSet(T) = \{term_1, term_2, \dots term_n\}$$
Where "n" is the number of terms composing the title "T."
- $\subset$ describes the relation of inclusion between titles. We note: $T_2 \subset T_1$ to indicate that the title "T_2" is a subtitle of the title "T_1."
- "Ttree(A)" is a tree that describes the hierarchy of titles in each document "A." It is defined by identifying the "root," the "nodes" and the "edges" of this tree.
- Root: title of the document A.
- Nodes: The set of titles and sub-titles belonging to the document A.
- Edges: each edge (T_1, T_2) corresponds to a relation of inclusion between two titles belonging to the same document A, i.e. $T_2 \subset T_1$.

The structure analysis method is composed of three phases: title pattern detection, three-level analysis and two-level analysis. These methods are described in the following sections.

4.1 Title pattern detection

To extract non-taxonomic relations, we propose analyzing the syntactic structure of the titles. For that, we manually define a syntactic pattern that determines the two arguments of each extracted relation as well as the corresponding label.

Principle: if there are two titles "T_1" and "T_2" belonging to the same article such that:

- $T_2 \subset T_1$
- T_2 is composed by a single noun phrase "Term""
- "T_1" has the syntactic structure:
  "S_{lab} prep $Term_1$, $Term_2$, … and $Term_n$"
  where "prep" is a preposition belonging to the set of prepositions $PrepSet = \{$"de,""de l′,""de la"$\}$ ($PrepSet = \{of\}$).

Then, "n" relations of the form ($Term_i$, Term', S_{lab}), for $1 \leq i \leq n$, are extracted.

The extracted relations are classified as taxonomic or non-taxonomic according to the value of "S_{lab}." Thus, if "S_{lab}" is equal to "type" or "kind," the relation is considered taxonomic. Otherwise, "n" non-taxonomic relations, having "S_{lab}" as label, are defined. The set of labels generated in this step is called "LabSet."

The title pattern detection step is presented in Algorithm 1.

Example:

Let us consider two titles "T_1" and "T_2" such that $T_2 \subset T_1$

T_1: Tool of medical biology // Outil de la biologie médicale

T_2: Macroscopic examination // Examen macroscopique

The title "T_2" is composed of a single noun phrase that is "macroscopic examination." So, $|Termset(T_2)|=1$. Moreover, the title "T_1" is of the form "term" "prep" "$term_1$" where "term" corresponds to "tool" and "$term_1$" corresponds to "medical biology." By applying the proposed approach, we obtain one non-taxonomic relation "R" = (medical biology, macroscopic examination, tool). This relation has as label "tool" and is valid and relevant, in the medical field, since macroscopic examination is a tool of medical biology.

4.2 Three-level analysis

The purpose of the "three-level analysis" is to retrieve labeled non-taxonomic relations between domain concepts by analyzing three levels of the title hierarchy in each document (Algorithm 2). The three-level analysis step is performed by analyzing the titles "T_1," "T_2" and "T_3" appearing in three nested levels of the title hierarchy and relies on the following heuristic:

Let "T_1," "T_2" and "T_3" be titles composed of one noun phrase with $T_3 \subset T_2 \subset T_1$. If "$T_2$" is a label ($T_2 \in$ LabSet), then a non-taxonomic relation "R" labeled by

Algorithm 1 *Title Pattern Detection*

Inputs: {*Edges*, *TR*, *NTR*, prepSet}
Outputs: {*RT*, *RNT*, *LabSet*}
LabSet= Ø
For each $u_i \in Edges$ \\ u_i is an edge of the form (T_1, T_2) where $T_2 \subset T_1$.
{
  Extract_args(u_i, T_1, T_2) //Extraction of the arguments of u_i
  If(($T_1 = S_{lab}$ prep S_1, S_2,…S_n) && (prep $\in$ prepSet) &&(|Termset(T_2)|=1))
    //Checking the syntactic structure of T_1 and the number of noun phrases composing T_2
  Then
  {
    If(S_{lab} = "type") //if the label of the relation is "type"
    Then For j from 1 to n **do**
      $TR = TR \cup (S_i, Termset(T_2))$ //Extraction of a taxonomic relation
    Else For j from 1 to n **do**
    {
      $NTR = NTR \cup (S_i, Termset(T_2), S_{lab})$ //Extraction of a non-taxonomic relation
      $LabSet = LabSet \cup \{S_{lab}\}$ //Extraction of the label
    }
  }
}

Algorithm 2 *Three-level analysis*

```
Inputs: {Edges, TR, NTR, LabSet}
Outputs: {RT, RNT}
For each u_i ∈ Edges \\ u_i is an edge of the form (T1, T2) where T2⊂T1.
{
    Extract_args(u_i, T1, T2) //Extraction of the arguments of u_i
    If(∃u' ∈ Edges / u'= (T2, T3)) //T3⊂T2
    Then If((|Termset(T1)|= |Termset(T2)|= |Termset(T3)|=1) && (Termset(T2) ∈ LabSet))
        //Checking the titles' length and the nature of the noun phrase composing T2
        Then
            {
                If(Termset(T2) = "type")  //If the relation label is "type"
                Then        TR= TR∪ (Termset(T1),Termset(T3))
                //A taxonomic relation linking the terms belonging to the titles of higher and lower levels
                //is extracted and added to the set of Taxonomic Relations TR
                Else NTR= NTR∪ (Termset(T1),Termset(T3), Termset(T2))
                //A non-taxonomic relation that has as label the noun phrase belonging to the title of the
                //medium level is extracted and added to the set of Taxonomic Relations NTR
            }
}
```

"T_2" and having as arguments Termset("T_1") and Termset("T_3") is extracted:
R = (Termset(T_1), Termset(T_3), Termset(T_2))

In the same manner as the title pattern detection approach, we consider a relation that has "type" as label to be taxonomic.

Example:
Let us consider the three nested titles "T_1," "T_2" and "T_3" extracted from a Wikipedia document where $T_3 \subset T_2 \subset T_1$:
T_1: Bronchioliti
T_2: Treatment
T_3: Physiotherapy

"T_1," "T_2" and "T_3" are composed of a single noun phrase. Moreover, the syntagm composing the title "T_2" constitutes a label in the medical field (Termset(T_2) ∈ LabSet). A non-taxonomic relation R = (bronchioliti, physiotherapy, treatement) linking bronchioliti and physiotherapy is then generated. This relation is labeled "treatment," which means that the physiotherapy is a treatment of bronchiolitis.

4.3 Two-level analysis

We propose, also, analyzing the titles belonging to two nested levels and composed of one noun phrase in order to see whether a relation could be generated.

Let T_1 (T_1= S_1) and T_2 (T_2= S_2) be two titles composed of one syntagm such that $T_2 \subset T_1$. Depending on the nature of the syntagms "S_1" and "S_2," the extracted relation is classified as either taxonomic or non-taxonomic. Indeed, if one of the two syntagms is a label already found then a non-taxonomic relation is generated otherwise a taxonomic relation is defined. Therefore, if:

1. Neither "S_1" nor "S_2" is a label ($S_1 \notin$LabSet and $S_2 \notin$LabSet) then a relation R = (S_1, S_2) is extracted. The expert has the task of validating the type of the extracted relation. If the domain expert considers the relation non-taxonomic, he performs the task of giving the label of the relation.

Example:
We extract from a Wikipedia article the two nested titles:
T_1= tooth decay // carie dentaire
T_2= devitalization // dévitalisation
When applying the two-level process, we obtain the relation (tooth decay, devitalization). We note that neither "tooth decay" nor "devitalization" is a label in the medical field. To evaluate the relation, we call on a domain expert. The latter fixes the relation as non-taxonomic and gives the label "treatment" to describe the relation between arguments. Other relations such as (cholera, media) are extracted by applying the two-level analysis on other Wikipedia articles but are rejected by the domain expert.

2. "S_1" is a label already found ($S_1 \in$ LabSet), then no relation is generated as the corresponding relation was already extracted by the three-level analysis method.
3. "S_2" is a label ($S_2 \in$ LabSet), then a non-taxonomic relation R = (S_1, X, S_2) labelled by "S_2" and having as first argument "S_1" is generated. To define the second argument of "R," we propose analyzing the paragraph that follows the title "T_2" in order to extract the missing argument "X." For that, we seek in this paragraph terms written with accentuated styles (bold, italic, etc.) and expressions corresponding to specific patterns. Indeed, pattern-matching methods are proven to be efficient when the relations labels are determined.

4.3.1. Patterns

To extract the missing argument "X" of the relation R = (S_1, X, S_2), where "S_1" is the syntagm composing a title "T_1" and "S_2" the syntagm composing a title "T_2" with $T_2 \subset T_1$ and $S_2 \in$ LabSet, we define the two patterns "P_1" and "P_2"

"P_1" is of the form: S_2 (adj) (of S_1) is A // S_2 (adj) (de S_1) est A
"P_2" is of the form: A is S_2 (adj) (of S_1) // A est S_2 (adj) (de S_1)

where (adj) and (of S_1) are optional.

We search in the paragraph following the title "T_2" expressions that have as syntactic form "P_1" or "P_2." For each extracted expression matching the pattern "P_1" or "P_2," a non-taxonomic relation of the form (S_1, A, S_2) is then generated.

Example:

Let consider the two nested titles "T_1" and "T_2" where T_1 = tuberculosis (tuberculose) and T_2 = treatment (traitement). As "treatment" is a label, the two-level analysis will generate a non-taxonomic relation R = (Tuberculose, X, treatment). To extract the missing argument X, we analyze the paragraph following the title "T_2" searching for an expression matching the defined patterns "P_1" and "P_2."

T_1= tuberculosis // tuberculose
T_2= treatment // traitement

… Rifadine is a treatment that can be used to treat tuberculosis of bone // … La rifadine est un traitement qui peut être utilisé pour traiter la tuberculose osseuse

We remark that the sentence "rifadine is a treatment" (rifadine est un traitement) matches the pattern P_2 (A is S_2), so the relation R= (tuberculosis, rifadine, treatment) is extracted.

4.3.2. Styles

In addition to patterns, we are interested in the styles used and links that occur in the paragraph that follows the subtitle. Common formatting (bold, italic, etc.) is usually used to highlight words that are assumed to be important in the context in which they appear. Because it is expected that the paragraph contains the desired argument, this one probably appears among the terms with specific styles.

In order to minimize errors and extract only interesting relationships, we propose a restriction of the length of the textual content that appears with a particular style. Thus, we are interested only in textual content composed by a maximum of two noun phrases. Each extracted noun phrase is considered as the second argument of the relation.

Example:

As an example, we consider a Wikipedia article dealing with "cancer." Among the subtitles of the title "cancer," we focus on "treatments." When applying the two-level process, we obtain the relation (cancer, treatment). Since "treatment" is a label, the relation is considered non-taxonomic. We focus on texts written with different styles in order to extract the second argument. Several relevant relationships are extracted, which we cite as examples:

(cancer, chemotherapy, treatment) //
(cancer, chimiothérapie, traitement)
(cancer, radiotherapy, treatment) //
(cancer, radiothérapie, traitement)
(cancer, ultrasound, treatment) //
(cancer, ultrason, traitement)

In summary, the two-level analysis is formalized in Algorithm 3.

5.0 Ontology refinement

Once the concepts and relations are extracted, we proceed to enrich our ontology by discovering new concepts and new relations. We have proposed previously (Zarrad, Doggaz and Zagrouba 2013) an enrichment method that extracts new concepts and new relationships and integrates them into the obtained ontology. A refinement of the obtained relations is then performed in order to keep only relevant relations. The refinement process is performed by checking three types of relations: transitive relations, conflictual relations and inverted relations.

5.1 Transitive relations

In this paragraph, we aim to eliminate redundant relations by referring to the transitive property. To do this, we focus on taxonomic relations already extracted. Among these re-

Algorithm 3 *Two-level analysis*

Inputs: {*Edges, TR, NTR, LabSet*}
Outputs: {*RT, RNT*}
For each $u_i \in$ Edges \\u_i is an edge of the form (T_1, T_2) where $T_2 \sqsubset T_1$.
{
Extract_args(u_i, T_1, T_2) //Extraction of the arguments of u_i
If(Termset(T_1)|=Termset(T_2)|=1)&&(Termset(T_1) $\notin$ *LabSet*))
//Checking the number of noun phrases composing T_1 and T_2 as well as the nature of noun phrase
//composing T_1
Then If(Termset(T_2) $\in$ *LabSet*) //Checking the nature of noun phrase composing T_2
Then
{
$NTR = NTR \cup$ Search_Pattern (Parag(T_2), P_1, P_2)
//Search_Pattern is a function that analyzes the paragraph following the title T_2 by
//seeking expressions of the form P_1 or P_2 and returns non-taxonomic relations
$NTR = NTR \cup$ Search_Style (Parag(T_2))
//Search_Style is a function that analyzes the paragraph following the title T_2 by seeking
//noun phrases having accentuated styles (bold, italic, etc) and returns non-taxonomic
relations

}
Else $TR = TR \cup$ (Termset(T_1), Termset(T_2))
//A taxonomic relation between the two noun phrases composing T_1 and T_2 is extracted and added
//to the set of Taxonomic Relations TR

lations, we check whether there are relations which can be deduced using the transitive property:

Let $R_1 = (S_1, S_2)$, $R_2 = (S_3, S_4)$ and $R_3 = (S_1, S_4)$ be three taxonomic relations. If the noun phrase $S_2 = S_3$ then the relation "R_3" is considered as redundant and is removed since it can be deduced from "R_1" and "R_2."

Example:
R_1= (disease, genetic disease) //
(maladie, maladie génétique)
R_2= (genetic disease, cancer) //
(maladie génétique, cancer)
R_3= (disease, cancer) //
(maladie, cancer)

According to the transitivity principle the relation "R_3" is then removed. Indeed, "R_3" can be deduced from "R_1" and "R_2."

5.2 Conflictual relations

Two relations "R_1" and "R_2" are considered conflictual if they have the same arguments but one is taxonomic while the other is non-taxonomic. Hence, the two relations $R_1 = (term_1, term_2)$ and $R_2 = (term_1, term_2, term_{lab})$ are conflictual and cannot coexist together. To resolve this problem, we favor the taxonomic relations that we consider more relevant than those that are non-taxonomic. Indeed, the methods used for taxonomic relation extraction are more reliable than those used for non-taxonomic relation discovery. Actually, when applying hierarchy analysis, especially title pattern detection and three-level analysis, we focus only on the hierarchy of the titles and not on text content. This could generate non-taxonomic instead of taxonomic relations. For this reason, we favor taxonomic relations over those that are non-taxonomic, and we consider that taxonomic relations have higher priority.

As an illustrative example, let us consider the following nested titles appearing in a Wikipedia article:

T_{gf}: Dyspnea "Dyspnée"
T_f : Causes "Causes "
T_{s1}: Chronic dyspnea "Dyspnée chronique"
T_{s2}: Acute dyspnea "Dyspnée aigüe"

When applying the three-level analysis method, we extract the two relations "R_1" and "R_2:"

R_1 = (dyspnea, chronic dyspnea, cause)
R_2 = (dyspnea, Acute dyspnea, cause)

We notice here that the two retrieved relations are invalid and that the links between "dyspnea" and "chronic dyspnea" on the one hand and "dyspnea" and "acute dyspnea" on the other hand are hierarchical. Indeed, the two subtitles "chronic dyspnea" and "acute dyspnea" are used by the article's author not as causes of the disease but in order to detail the causes of dyspnea in chronic and acute cases. The causes of dyspnea are then described in the paragraphs that follow each subtitle. Thus, the two relations "R_1" and "R_2" are not correct and have to be removed. Let us assume, also, that we have a taxonomic relation R = (dyspnea, chronic dyspnea). Since the two relations "R_1" and "R" are conflictual, the relation "R_1" will be removed and only the relation "R" is retained. The same principle can be applied to some examples of relations extracted when applying the title pattern detection step.

5.3 Inverted relations

The last step of relation refinement focuses on taxonomic relations having inverted arguments. Let "R_1" and "R_2" be two taxonomic relations such that:

R_1 = ($term_1$, $term_2$)
R_2 = ($term_2$, $term_1$)

In other words, "$term_2$" is considered as a sub-concept of "$term_1$" and "$term_1$" is also considered as a sub-concept of "$term_2$." This double inclusion cannot be retained. Only the valid relation is kept. The other is removed from the set of taxonomic relations.

In order to select the valid relation, we classify the taxonomic relations in four classes according to the method used for their selection:

- C_1: The relations extracted using pattern's recognition method;
- C_2: The relations extracted using title pattern detection and three-level analysis methods and that have the term "type" as label;
- C_3: The relations selected by syntactic analysis;
- C_4: The relations selected by the two-level method then validated and classified as taxonomic ones by the domain expert.

In order to resolve the conflict and choose the relevant relation, we focus on method reliability. We check the relevance of each relation by considering the method that has been used to extract it. Thus, we assign the weight α_i for each class "C_i" reflecting the importance given to the method used. For each extracted relation "R," we assign a binary value "$T_i(R)$" such as $T_i(R)$ is equal to one if "R" belongs to "C_i," otherwise $T_i(R) = 0$. We then compute the relation weight "T(R):"

$$T(R) = \sum_{i=1}^{4} \alpha_i * T_i(R) \qquad \text{(eq. 3)}$$

with:

$$\begin{cases} \sum_{i=1}^{4} \alpha_i = 1 \\ 0 < \alpha_1 < \alpha_2 < \alpha_3 < \alpha_4 < 1 \\ \alpha_1 + \alpha_2 + \alpha_3 < \alpha_4 \\ \alpha_1 + \alpha_2 < \alpha_3 \end{cases} \qquad \text{(eq. 4)}$$

The greatest weight is assigned to relations identified by the domain expert (α_4). It is interesting to note that this weight is chosen such that a relation validated by a domain expert eliminates any other relation. Hence, the constraint $\alpha_1 + \alpha_2 + \alpha_3 < \alpha_4$ of eq. 4. We note that a relation "R" can belong to different classes. The syntactic analysis also gives interesting results. Indeed, the adjective makes a term more specific, so that a term "T_1" composed by a term "T_2" followed by an adjective is usually a sub-concept of "T_2." Thus, the relations belonging to the class "C_3" have the second priority. The third priority is assigned to the class "C_2." These relations extracted when analyzing the titles' hierarchy and having "type" as label can be considered taxonomic. We give the last priority to the class "C_1," which is the least efficient.

The coefficient constraints ensure that two inverted relations belonging to two different classes could not have the same weight. The relation having the highest weight is retained, the other one is removed.

6.0 Experimental study

To evaluate the proposed ontology building approach, we are interested in French Wikipedia articles related to the medical field. This field is terminologically rich and has been the focus of several research projects (Sawsaa and Lu 2014; Tang et al. 2017). Our corpus is composed of 1,222 Wikipedia medical articles. 189,900 CT are extracted from the corpus where:

- 110,929 CT belong to class "Cl_1." Some examples of CT belonging to "Cl_1" are: lung, cell, tumor, disease, etc.
- 78,971 CT belong to class "Cl_2." Some examples of CT of the class "Cl_2" are: genetic disease, human body, immune system, etc.

Then, we apply the relation extraction approach, we obtain as result 17,114 relations:

– 3,765 taxonomic relations obtained by performing a syntactic analysis of the corpus: 1,854 before enrichment and 1,911 when applying enrichment step.
– 3,463 taxonomic relations and 5,722 non-taxonomic relations extracted when projecting the lexical-syntactic patterns on the corpus.
– 4,164 relations (taxonomic and non-taxonomic) obtained by exploring the hierarchy of titles in each document.

The evaluation of the ontology relations is first conducted by a domain specialist using a random sample of 500 relations. He checks to see whether each retrieved relationship is relevant to the medical field. Then, we compute the precision and recall values. We use the same definitions of precision and recall known in information retrieval (IR) (Powers 2011). The second step of ontology evaluation consists of comparing our ontology to other existing ontologies and knowledge repositories. In order to compare ontologies, we were inspired by the measures proposed in Brank et al. (2005). The authors propose two measures: precision and recall. We note that these measures and those known in IR (Powers 2011) are not the same. In Brank et al., precision is the percentage of the ontology lexical entries that also appear in the gold standard, relative to the total number of ontology words. Recall is the percentage of the gold standard lexical entries that also appear as concept identifiers in the ontology, relative to the total number of gold standard lexical entries. We apply these measures not only for lexical entities (concepts) but also for the evaluation of ontological relations. A way to achieve more tolerant matching criteria is to consider all subclasses of each lexical entry and all their subclasses belonging to lower levels. Then, instead of testing whether the extracted taxonomic relation ("C_1," "C_2") appears also in the gold standard, we propose to check whether the concept "C_2" belongs to the whole set of subclasses of "C_1" that we note Sub(C_1). The same idea is applied for recall measure. To distinguish these measures from known precision and recall measures, we call them P_{Gold} and R_{Gold}.

The precision values given by relation extraction methods are given in Table 1. To compute these values, we use simples composed by 500 relations. As we can note, the syntactic analysis and the three steps of title hierarchy analysis give the best results. The pattern recognition approach extracts at the same time taxonomic and non-taxonomic relations. Nevertheless, this method gives better results for taxonomic relation discovery. We note also that the number of extracted relations is relatively high. This is due to the long list of patterns used.

When exploring the hierarchy of the titles in each document, we extract not only taxonomic relations but also non-taxonomic. The resulting set of non-taxonomic relations is automatically labelled without the intervention of the expert. The corresponding precision values are interesting and validate our intuition. We obtain as recall values 54.55% for taxonomic relations extraction and 73.42% for non-taxonomic. The corresponding F-measure values are 58.56% and 72.73%. Before checking the relations arguments, the two-level analysis step gives a total number of 3,018 relations. We obtain 739 relations with a label as second argument. We scan the paragraph that follows each subtitle by checking patterns and styles used in order to extract the missing argument of each relation.

In Table 2, we present a description of the extracted relations and the corresponding precision values when applying the two-level analysis step. We note that 2,077 new non-taxonomic relations are retrieved. Among them, 434 relations are rectified by the expert by modifying a relation argument or label.

Afterwards, we apply a refinement to the set of all extracted relations in order to eliminate redundancy and remove invalid relations. The resulting set of relations is composed of 5,679 taxonomic relations and 8,065 non-taxonomic. The number of relations that are removed referring to the property of transitivity is thirty-six. The analysis of conflictual relations gives forty-one pairs of relations (taxonomic and non-taxonomic). We keep only taxonomic relations that correspond to a precision value of 85.37%. In the last step of relation refinement, we analyze relations having inverted arguments. Four pairs of taxonomic relations are retrieved. We choose the values $\{\alpha_4 = 0.6, \alpha_3 = 0.25, \alpha_2 = 0.1, \alpha_1 = 0.05\}$ that verify equ. 4. By giving priority to each class of relations referring to the methods used, we keep only those that are relevant. This step gives 100% as precision value.

To compare our results to other existing ontologies, we chose those dealing with the same field and the same language. We used the French version of *Medical Subject Headings* (*MeSH*) and the category system in Wikipedia. *MeSH* is a taxonomic hierarchy of medical and biological terms suggested by the U. S. National Library of Medicine (Soualmia et al. 2004). This thesaurus contains about 27,582 classes that represent the basic concepts in the medicine field. The category system of Wikipedia corresponds to a thematic classification system of Wikipedia articles that appears at the bottom of each article page. It was used for evaluation in other works (Paukkeri et al. 2012). We focus on the French category "medecine" which contains twenty-eight subcategories. Each of these subcategories gives rise to other subcategories, etc.

When computing the P_{gold} and the R_{gold} on a random sample of about 100 taxonomic relations extracted by our system, we conclude that 31.68% of these relations appear also in *MeSH* and 28.71% appear in Wikipedia categories. To compute recall, which corresponds to the percentage

Method		Extracted relations			
		Taxonomic relations	Precision	Non-taxonomic relations	Precision
Syntactic analysis		1854	84%	×	×
Patterns matching		3463	60.31%	5722	41.22%
Titles' Hierarchy Analysis	Title pattern detection	20	100%	96	64.58%
	Three-level Analysis	×	×	191	90.05%
	Two-level Analysis	1780	62.8%	2077	78,75%

Table 1. Evaluation of the proposed methods.

	Number of extracted relations	Correct relations (Precision values)	Rectified relations
Patterns	14	71.43%	0
Styles	2063	57.8%	21%

Table 2. Evaluation of two-level analysis step.

Reference Measures Repositories	P_{gold}	R_{gold}
MeSH	31.68%	43.56%
Wikipedia Categories	28.71%	53.06%

Table 3. Ontologies comparison.

of the gold standard relations that also appear in our ontology, we use a sample of similar size. The results are given in Table 3.

We notice that our ontology gives more interesting results than the two referenced systems. Moreover, *MeSH* and the category system of Wikipedia give mainly taxonomic relations. However, we notice that many relations given by these two repositories are not taxonomic. For example, when exploring the medical category, we find that, on the one hand, the category "discipline of chemistry" (discipline de la chimie) is a subcategory of "chemistry" (chimie). On the other hand, the category "biochemistry" (biochimie) is a subcategory of "discipline of chemistry." The first relation is not a taxonomic relation. These two relations could be replaced by a single non-taxonomic relation labelled "discipline" linking the two concepts "chemistry" and "biochemistry." Many other examples of such relations are found either in *MeSH* or in the categories of Wikipedia. In our ontology, this problem does not arise since we extract directly the non-taxonomic relation with the associated label.

7.0 Conclusion

In this paper, we have presented an extended version of our approach that deals with building ontology using document structure. The proposed methods focus mainly on the title hierarchy to extract semantic relations. These methods extract not only taxonomic relations but also non-taxonomic relations. Most of these relations are labelled automatically. Relations discovery is performed in three steps: title-pattern detection, three-level analysis and two-level analysis. The latter two approaches have the advantage of not being language specific. In three-level analysis step, we have defined several assumptions in order to detect the relation type, the relation arguments and the corresponding label in case of non-taxonomic relation. The methods proposed for relations discovery give a large number of relations. We propose to apply a refinement to the set of all extracted relations in order to remove redundant and invalid relations keeping only relevant ones. The refinement process is performed by applying transitive property, checking the nature of the relations and analyzing taxonomic relations having inverted arguments. A comparison with a gold standard is finally performed in order to evaluate the generated ontology. The proposed approach is evaluated on French Wikipedia articles related to the medical field. Nevertheless, we note that this approach could be applied on other fields to extract taxonomic and non-taxonomic relations. Indeed, in the proposed approach, the relations labels are extracted automatically. Therefore, any kind of relations could be extracted. We note that the relations arguments are composed either of a "name" or a "name" followed by an "adjective." We are interested in these nominal groups since they are those most used in technical texts.

As future work, it might be interesting to extend the extraction of concepts to more complex noun phrases. They may include, in addition to the "noun," a preposition, a proper noun, an adverb, etc. We mention as examples "Ministry of Finance," "Monte Carlo analysis" and "sexually transmitted diseases." Such CT are semantically rich and often appear in technical corpora. Moreover, a refinement of the set of concepts having more than one hypernym can be conducted. In other words, if "$term_1$" is, in the same time, a sub-concept of "$term_2$" and a sub-concept of "$term_3$," it would be interesting to select the relevant relation and remove the invalid one. We have noted that, in

some cases, there is a taxonomic relation between "$term_2$" and "$term_3$." We aim also to apply a refinement to non-taxonomic relations that have the same arguments but have different labels. Indeed, if "R_1" is of the form ($term_1$, $term_2$, lab_1) and "R_2" is of the form ($term_1$, $term_2$, lab_2), it would be judicious to check the link between the two terms "$term_1$" and "$term_2$" and specify the correct label.

Furthermore, we aim to increase the corpus size in order to improve precision and recall values. Considering the fact that the field on which we work is wide and contains a huge number of concepts, we will probably obtain more interesting results when the number of documents in the corpus is high. For the time being, evaluation is performed by only one domain expert. A deeper evaluation realized by several domain experts is planned for future work.

References

Abd El-atey, Amira, Sherif El-etriby and Arabi S. kishk. 2012. "Semantic Data Extraction from Infobox Wikipedia Template." *International Journal of Computer Applications* 40, no. 17: 18-23.

Almeida Campos, Maria Luiza de, and Hagar Espanha Gomes. 2017. "Ontology: Several Theories on the Representation of Knowledge Domains." *Knowledge Organization* 44:178-86.

Arnold, Patrick and Ehard Rahm. 2014. "Extracting Semantic Concept Relations from Wikipedia." In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14) … Thessaloniki, Greece, June 02-04 2014*. New York: ACM, 26-36. doi:10.1145/2611040.2611079

Aussenac-Gilles, Nathalie and Mouna Kamel. 2009. "Ontology Learning by Analyzing XML Document Structure and Content." In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development 6-8 October 2009, Madeira, Portugal*, 159-65. doi: 10.5220/0002293301590165

Barbu, Eduard and Poesio Massimo. 2009. "Unsupervised Knowledge Extraction for Taxonomies of Concepts from Wikipedia." In *Proceedings of the International Conference in Recent Advances in Natural Language Processing 14-16 September 2009*, *Borovets, Bulgaria*, 28-32.

Barrière, Caroline. 2016. "Pattern-Based Relation Extraction." Chap. 11 in *Natural Language Understanding in a Semantic Web Context*, 205-29. Switzerland: Springer.

Brank, Janez, Marko Grobelnik, and Mladenic Dunja. 2005. "A Survey of Ontology Evaluation Techniques." In *Proceedings of the 8th International Multi-conference Information Society 11-17 October 2005, Ljubljana, Slovenia,* 166-170.

Cartier, Emmanuel. 2015. "Extraction automatique de relations sémantiques dans les définitions: approche hybride, construction d'un corpus de relations sémantiques pour le français." In *Conférence annuelle Traitement Automatique des Langues Naturelles, Jun 2015, Caen, France. Traitement Automatique des Langues, conférence internationale, 2015*. Hal ID: halshs-01412736. https://halshs.archives-ouvertes.fr/halshs-01412736/document

Filipiak, Dominik and Ławrynowicz Agnieska. 2014. "Generating Semantic Media Wiki Content from Domain Ontologies." In SWCS'14 Proceedings of the Third International Conference on Semantic Web Collaborative Spaces, Riva del Garda, Italy. Aachen, Germany: CEUR, 68-76.

Harris, Zellig. 1954. "Distributional structure." *Word* 10: 146–62.

Hazman, Maryam, Samhaa R. El-Beltagy, and Ahmed Rafea. 2011. "A Survey of Ontology Learning Approaches." *International Journal of Computer Applications* 22, no. 9: 36-43.

Hearst. Marti. 1992. "Automatic Acquisition of Hyponyms from Large Text Corpora." In *Proceedings of the 14th Conference on Computational Linguistics Nantes, France*, *23-28 August 1992*. Stroudsburg, PA: Association for Computational Linguistics, 539-45. doi: 10.3115/992133.992154

Joachims, Thorsten. 1997. "A Probabilistic Analysis of the Rocchio Algorithm with Tfidf for Text Categorization." In *Proceeding of the 14th International Conference on Machine Learning 8-12 July 1997, Nashville, Tennessee, USA*. 143-51.

Jacques, Marie-Paule and Josette Rebeyrolle. 2006. "Titres et structuration des documents." In *Actes International Symposium: Discourse and Document 15-16 Juin 2006, Caen, France,* 1-12.

Kermanidis, Katia and Nikos Fakotakis. 2007. "One-sided Sampling for Learning Taxonomic Relations in the Modern Greek Economic Domain." In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, 29-31 October 2007, Patras, Greece*. Washington, D.C.: IEEE, 354-61. doi: 10.1109/ICTAI.2007.196

Loffler-Laurian, Anne-Marie. 1994." Réflexions sur la métaphore dans les discours scientifiques de vulgarisation." *Langue française* 101: 72-79.

Marshman Elizabeth. 2008. "Expressions of Uncertainty in Candidate Knowledge-Rich Contexts: A Comparison in English and French Specialized Texts." *Terminology* 14: 124-51.

Marshman, Elizabeth, Tricia Morgan, and Ingrid Meyer. 2002. "French Patterns for Expressing Concept Relations." *Terminology* 8: 1-29.

Medelyan, Olena, Steve Manion, Jeen Broekstra, Anna Divoli, Anna-Lan Huang, and Ian H. Witten. 2013. "Constructing a Focused Taxonomy from a Document Collection." In *The Semantic Web: Semantics and Big Dat: 10th International Conference, ESWC 2013, Montpellier,*

France, May 26-30, 2013. Proceedings, ed. Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph. Lecture Notes in Computer Science 7882. Berlin: Springer Berlin Heidelberg, 367-81.

Morin, Emmanuel. 1999. "Using Lexico-Syntactic Patterns to Extract Semantic Relations between terms from Technical Corpus." In *TKE '99 Terminology and Knowledge Engineering: Proceedings, Fifth International Congress on Terminology and Knowledge Engineering, 23-27 August 1999, Innsbruck, Austria*, ed. Peter Sandrini. Vienna: TermNet, 268-78.

Morlane-Hondere, François and Fabre Cécile. 2010. "L'antonymie observée avec des méthodes de TAL: une relation à la fois syntagmatique et paradigmatique?" In *Proceedings of TALN 19–23 July 2010, Montréal, Canada*. 6-13.

Panchenko, Alexander, Stefano Faralli, Eugen Ruppert, Steffen Remus, Hubert Naets, Cedrick Fairon, Simone Paolo Ponzetto and Chris Biemann. 2016. "TAXI: A Taxonomy Induction Method Based on Lexicosyntactic Patterns, Substrings and Focused Crawling." In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. San Dieg: Assn. for Computational Linguistics, 1320-27.

Paukkeri, Mari-Sanna, Alberto Pérez García-Plaza, Víctor Fresno, Raquel Martínez Unanue, and Timo Honkela. 2012. "Learning a Taxonomy from a Set of Text Documents." *Applied Soft Computing* 12: 1138-48.

Pieterse, Vreda and Derrick G. Kourie. 2014. "Lists, Taxonomies, Lattices, Thesauri and Ontologies: Paving a Pathway Through a Terminological Jungle." *Knowledge Organization* 41: 217-29.

Ponzetto, Simono Paolo and Michael Strube. 2007. "Deriving a Large Scale Taxonomy from Wikipedia." In *Proceedings of the Twenty-second AAAI Conference on Artificial Intelligence: 22-26 July 2007, Vancouver, British Columbia, Canada*. Menlo Park, CA: AAAI Press, 1440-45.

Powers, D M. W. 2011. "Evaluation: From Precision, Recall and F-Factor to Roc, Informedness, Markedness and Correlation." *Journal of Machine Learning Technologies* 2: 37-63.

Punuru, Janardhana and, Jianhua Chen. 2012. "Learning Non-Taxonomical Semantic Relations from Domain Texts." *Journal of Intelligent Information Systems* 38: 191-207.

Rios-Alvarado, Ana B., Ivan Lopez-Arevalo and Victor Sosa-Sosa. 2013. "Learning Concept Hierarchies from Textual Resources for Ontologies Construction." *Expert Systems with Applications* 40: 5907-15.

Sanchez, David and Antonio Moreno. 2008. "Learning Non-Taxonomic Relationships from Web Documents for Domain Ontology Construction." *Data and Knowledge Engineering* 64: 600–23.

Sawsaa, Ahlam F. and Joan Lu. 2014. "Using Natural Language Programming (NLP) Technology to Model Domain Ontology OTO by Extracting Occupational Therapy Concepts." *Knowledge Organization* 41: 452-64.

Shen, Minxin, Duen-Ren Liu, Yu-Siang Huang. 2012. "Extracting Semantic Relations to Enrich Domain Ontologies." *Journal of Intelligent Information Systems* 39: 749-61.

Shamsfard, Mehrnoush and Ahmad-Abdollahzadeh Barforoush. 2003. "The State of the Art in Ontology Learning: A Framework for Comparison." *The Knowledge Engineering Review* 18: 293-316.

Schmid, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees." In *Proceedings of the Conference: International Conference on New Methods in Language Processing; (NeMLaP), September 14-16 1994 The University of Manchester Institute of Science and Technology Manchester United Kingdom*, ed. Daniel Jones. Manchester: Centre for Computational Linguistics, 44-49.

Snow, Rion, Daniel Jurafsky, and Andrew Ng. 2005. "Learning Syntactic Patterns for Automatic Hypernym Discovery." In *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference,* ed. Lawrence K. Saul, Yair Weiss and Léon Bottou. Cambridge, MA: MIT Press, 1297-1304.

Soualmia, Lina-Fatima, Christine Golbreich, and Stéfan-Jacques Darmoni. 2004. "Representing the MeSH in OWL: Towards a Semi-Automatic Migration." In *Proceedings of the KR 2004 Workshop on Formal Biomedical Knowledge Representation Whistler, BC, Canada, 1 June 2004*, ed. Udo Hahn, 81-87.

Suchanek, Fabian, Gjergji Kasneci, and Gerhard Weikum. 2008. "Yago: A large Ontology from Wikipedia and Wordnet." *Web Semantics: Science, Services and Agents on the World Wide Web* 6: 203-17.

Sumida, Asuka and Kentaro Torisawa. 2008. "Hacking Wikipedia for Hyponymy Relation Acquisition." In *Proceedings of the Third International Joint Conference on Natural Language Processing, January 7-12, 2008, Hyderabad, India* 2: 883–888. doi:10.1.1.178.2699

Tang, Xiao-Bo, Wei Wei, Guang-Chao Liu and Juan Zhu. 2017. "An Inference Model of Medical Insurance Fraud Detection: Based on Ontology and SWRL." *Knowledge Organization* 44: 86-94.

Weichselbraun, Albert, Gerhard Wohlgenannt, Arno, Scharl, Michael Granitzer, Thomas Neidhart, and Andreas Juffinger. 2009. "Discovery and Evaluation of Non-Taxonomic Relations in Domain Ontologies." *International Journal of Metadata, Semantics and Ontologies* 4: 212-22.

Wong, Wilson, Wei Liu, and Mohammed Bennamoun. 2012. "Ontology Learning from Text: A Look Back and into the Future." *ACM Computing Surveys* 44: 20. doi:10.1145/2333112.2333115

Wu, Fei and Daniel Weld. 2008. "Automatically Refining the Wikipedia Infobox Ontology." In *Proceedings of the*

17th International World Wide Web Conference 21-25 April 2008, Beijing, China, 635-44. doi:10.1145/1367497.1367583

Zarrad, Rim, Narjes Doggaz and Ezzeddine Zagrouba. 2012. "Toward a Taxonomy of Concepts using Web Documents Structure." In *iiWAS2012: Proceeding, The 14th International Conference on Information Integration and Web-based Applications & Services, December 3-5 2012, Bali, Indonesia*, ed. David Taniar et. al. New York: ACM, 303-12.

Zarrad, Rim, Narjes Doggaz and Ezzeddine Zagrouba. 2013. "Title-Based Approach to Relation Discovery From Wikipedia." In *Proceedings of the 5th International Conference on Knowledge Engineering and Ontology Development, Vilamoura, Algarve, Portugal, er 19-22 September 2013*, ed. Joaquim Filipe et al. [Setúbal]: SciTePress, 70-80.