ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. I, Issue 10, December 2013

Semi-Automatic Enrichment Approach of 'Domain Ontology' by using TALN Tools

Khalid TATANE[1], Brahim ER-RAHA[2], Sanaa MOUHIM[3], Chihab eddine CHERKAOUI[4]

Laboratory of Industrial and IT Engineering, National school of Applied Science, B.P: 1136/S Agadir-Morocco[1, 2]

Laboratory IRF-SIC, Faculty of Science Agadir, B.P: 37/S Agadir-Morocco [3, 4]

ABSTRACT: In most of researches, the semi-automatic construction of ontology using texts is generally based on textual medium, and in which the studied domain describes text content. In order to conceive semantically the richest and updated ontology, we suggest developing the classical methods of ontology construction (1), by taking into account the text "content" to construct first kernel of ontology, and (2) by enriching obtained ontology using external resources (public texts, controlled vocabulary of the same domain, specialized web pages, text documents, etc.). This paper describes how these different resources are analyzed and exploited.
We have experimented this approach on texts dealing with subjects related to tourism in morocco, and we have assessed the advantage drawn from one richest ontology of the domain (in comparison to a first realized ontology of the domain in the framework of the project GECO-WES[1]), and we have noted that the results have seen significant improvement.

Keywords: Domain ontology, Corpus, Normalization, TALN Tools, Candidate terms, Concept, Semantic relation, Formalization, Ontology enrichment.

I. INTRODUCTION

The construction and the populating of the ontology depend strongly on data extracted from the various information sources. In the manual mode, the analysts of the domain base themselves on classical techniques to collect information, such as discussions with the experts of the domain or by manual analysis of documents.
However, this manual processing of documents is extremely expensive in time and resources. The whole process also raises problems of productivity and quality. In the semi-automatic mode, ontology building using texts relies often on the process of text analysis, whether it is according to statistical approach: (Aguirre and al., [1]; Faatz and al., [2]; Parekh and al., [3]) or linguistics approach: (Nédellec and Zweigenbaum, [4]; Aussenac and al., [5]; Maedche, [6] Buitelaar and al., [7]). The processing tools of languages are used to analyze texts and extract semantic concepts and relations. In fact, a text is an important source of knowledge, which is constant and shared by practicing communities. Texts contain linguistic elements such as candidate terms, semantic classes and relations, which are very useful to the ontology construction of the domain.

In addition, texts are more available than experts of the domain that intervene at the modeling level. It is worth mentioning that the construction process cannot be fully automatic, because the results of extractors are noisy, and that necessitates a subjective judgment of the ontology specialist.

To be able to analyze a text and extract information, many platforms of automatic processing of Natural Languages exist. However, there is no standard methodology which constitutes a unifying sustained framework for semi-automatic construction of the ontology using textual data (expressed in natural language). Generally, a framework of four stages is common in most methodologies of ontology construction using text. These four stages are: 1) constitution of documents corpus, 2) linguistic analysis of the corpus, 3) normalization (Aussenac-Gilles, [8]; Cimiano & Volker, [9]), and 4) formalization of the ontology.

[1] GECO-WES: (**GE**stion de **CO**nnaissance et **WE**b **S**émantique), Knowledge Management and Semantic Web.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. I, Issue 10, December 2013

The methodology we propose in this paper is not different from the common one, but it is mainly characterized by rich and diversified nature of the resources used. We try also to define all TALN[2] tools, relying on reactions we got, in order to efficiently and subjectively come up with tools of different stages of domain ontology construction.

This work is conducted in the framework of GECO-WES project research, which deals with general issue of creation, development and maintenance of ontology for semantic web, so as to ameliorate information research in the domain of tourism in Morocco.

For these reasons, and in order to clarify our scientific reasoning, we divided this paper into four parts. The first one covers creation of domain corpus, its pre-processing and its normalization. The second part shed lights on linguistic analysis which consists of morpho-syntactic labeling, the candidate term extraction, and the semantic relations extraction. The third part deals with normalization analysis of the ontology by verifying and integrating extracted concepts and roles in base ontology. The fourth part is related to ontology formalization.

II. METHODOLOGICAL APPROACH OF SEMI-AUTOMATIC ENRICHMENT OF DOMAIN ONTOLOGY

The semi-automatic enrichment of the initial ontology consists in detecting, by using text: terms, concepts and semantic relations, to handle them, and to integrate them into ontology under a process of development, which is called kernel base (S.Mouhim, [10]).

In literature, many techniques and algorithms of automatic processing of language are used. In this part, we are interested in results exploitation of different types of software used in language processing, and which permit the extraction of candidate terms (Concepts) and semantic relations between them.
Generally speaking, there are four common stages in all methodologies of construction and enrichment of domain ontology by using text:

- **Constitution of corpus.**
- **Linguistic analysis of corpus.**
- **Normalization in concepts and in semantic relations.**
- **Formalization of the ontology.**

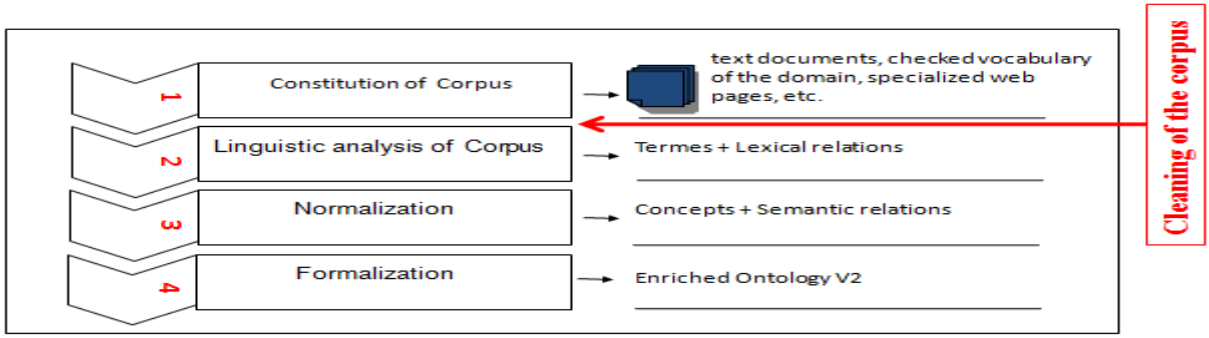

Fig. 1 Main stages of construction and enrichment of a domain ontology from text.

The methodology we propose is mainly characterized by a pre-processing stage, which is often ignored in most experiments of ontology construction by using text. In fact, the TALN tools are very sensitive to noise which may contain a textual document. The cleaning and repairing of corpus influence, as we will see later, the quality and quantity of extracted information. The second contribution consists of the integration and exploitation of results of different extractors of terms and relations. As matter of fact, all works that we have consulted use the results of one extractor of terms and relations. In this work, we suggest combining the results of two terms extractors Yatea and TermoStat. The idea we are defending is that one term candidate has more chance to be selected when it is extracted by many tools.

[2] TALN: Automatic Treatment of Natural Languages.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. I, Issue 10, December 2013

When setting up our approach, many important questions raised, to which we have tried to provide an answer. They are mainly about the choice of adequate corpus, the necessary number of documents to constitute a good corpus, TALN tools to choose, sorting and selecting candidate terms and relations, and finally the question of integrating results in the base ontology.

All conducted reactions during stages of the adopted methodology which uses textual medium are presented in the following parts.

A. CONSTITUTION OF CORPUS

A corpus is a set of texts considered to be a representative of a language, dialect or a part of language used in linguistic analysis. It is a key element in every text analysis reasoning. In fact, a corpus is made of several linguistic elements from which ontological and terminological resources are constructed. Consequently, selecting a text should be subject to some criteria.
First, and as we are mainly aiming to enrich a domain ontology, it is appropriate then to choose specialised texts reflecting, as much as possible, the domain tackled. Besides, the qualifications and expertise of the author are necessary. For terminological purposes, it is necessary to use written texts by experts of the domain.

Concerning the size of text, there is always no agreement on the appropriate size for a corpus. (Pearson, [11]) precises that the size of a specialised corpus is determined according to researcher intuition, character of included texts and their domain. We propose the average of one million words[3].

The chosen corpus should also guarantee a global heterogeneous coverage of the domain to take in consideration different points of view, like, for example, scientific magazine, textbook, newspapers, as it might be web sites, though their content is not always reliable.

Summing up, no matter what the used resource is, it should be chosen carefully so as to cover the global domain of knowledge. The developer of ontology can choose between use of the totality or a part of the resource, if not combining many resources of different types. There are no particular criteria to choose textual resources, and what is important is being aware of information relevance in such mediums.

In this paper, we have used official websites of some tourism organisations (S.Mouhim, [12]) so as to guarantee the reliability, quality and quantity of the information. The chosen web sites are of different types Like blogs, wikis, portals, etc. The aim is to interconnect many point of view on the studied domain (approach of Kilgarriff, [13] for corpus construction by collecting web pages).

The corpus constitution process generates a set of sources of knowledge divided as follows: 300 web sites pages in French belonging to 20 web sites of different nationalities (Morocco, France, USA, Canada, Belgium...), 10 documents PDF, 8 presentations Power Point, 10 tourism magazines, which gives a total of 1 115 922 words to be analyzed.

B. PRE-PROCESSING AND NORMALIZATION OF THE CORPUS

The pre-processing of the corpus and its normalization constitute a key stage in the automatic processing of languages. It is matter of decreasing, if not eliminating, lexical incoherence and syntactical ambiguities which a text contains. In fact, the text searching tools are sensitive to an important noise caused for example by typing error or errors due to conversions and passages from one format another.

In literature, the normalisation stage consists mainly of lemmatization of words and segmentation of sentence (Korenius et al., [14]; Zhang et al., [15]).

[3] The choice of the figure is justified according to (Pearson, 1998 [11]), by the fact that the corpus should reach a critical size permitting statistical accurate treatment. In fact, the more the size corpus augment, the more proven the results are.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. I, Issue 10, December 2013

- **Lemmatization:** refers to lexical analysis, which requires finding base form of an inflected word called lemma. This stage permits reducing numbers of forms to be considered, and then the augmentation of occurrences of every form in the corpus. For example, the French words: (est, etant, was reduced to the verb «être»).
- **Segmentation:** refers to automatic division of a text to smallest units by learning of sentences borders.

There are other types of pre-processing of texts. They are unfortunately less tackled or dealt with discreetly. They are especially the preprocessing that precedes lemmatization and segmentation of text stated in the works of (Grefenstette et al., [16]; Habert et al., [17]).

According to our experiment, a first processing of the corpus resides in the unification of formats HTML, PDF, PPT, etc. to raw text format.

To identify the incoherence contained in the corpus, we have experimented a labelling morpho-syntactical tool on tested documents. We have noted that there are confusions mainly due to characters encoding, to punctuations, to spelling mistakes, to figures, to measure units and to upper cases.

```
2    NUM @card@
500  NUM @card@
m    NOM <unknown>
en   PRP en
face     NOM face
nord     ADJ nord
"    PUN:cit "
mulet-ski    NOM <unknown>
"    PUN:cit "
par PRP par
mini-bus     NOM <unknown>
,    PUN ,
par PRP par
taxiscollectifs NOM <unknown>
ministere    NOM <unknown>
de   PRP de
l'economie   NOM <unknown>
```

Fig. 2 Example of incoherence detected after trial with a labeling morpho-syntactical tool due mainly to the lack of normalization and pre-processing stage in the corpus.

After the incoherence detected in the labelling morpho-syntactical stage, we propose to return to the creation stage of the corpus and to carry out a correction according to the following approach:

1) Resolution of detected problem in characters with accent by using a unified coding UTF8;
2) Correction of spelling mistakes due to typing errors or passage from one format to another;
3) Words with upper-case which do not occur in the beginning of sentence constitute a third type of incoherence. Lower-case formatting is necessary;
4) Punctuation processing except those representing compound words or delimiting a sentence, for example '-' and '.'. In fact, in the case of compound words '-' external dictionaries are used such as BDLEX (Pérennou, [18]). If a word is defined in the dictionary, no modification is to be made. However, in the opposite case, this punctuation is separated from text as in the example: Marrakech-Agadir which becomes after correction Marrakech – Agadir. In the same context, some words may contain full stops in the middle of a sentence (example: « ex. titre », e.g. résultat). These full stops are omitted to avoid any confusion with borders of sentences;
5) Processing of abbreviations and acronyms;
6) Marking of paragraphs borders by spaces;
7) Processing of figures, in this case, we converted numbers into their textual equivalents, for example, « 1 » becomes «FR: un | EN: one »;
8) The replacement of units symbols such as "m, kg, km²", to become « m » by « FR: mètre | EN: meter» and « ² » becomes « FR: carré | EN: cube».

Once the corpus is cleaned and normalized, the next stage is its linguistic analysis.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. I, Issue 10, December 2013

C. LINGUISTIC ANALYSIS

1) *Morpho-syntactical labelling*

Morpho-syntactical labelling of a text consists of identifying for every word its morpho-syntactical category, that is, grammatical category, gender, number, tense, etc. In the Web, there is many tools of labelling such as Cordial[4], Brill[5], Tree-Tagger[6], Multext[7], Genia-Tagger[8] , Lia-Tagg[9], Pos-Tagger[10].Each of these tools is based on different linguistic hypotheses.

Our purpose is neither carrying out a comparative study of different tools of morpho-syntactical labelling which exist in literature, nor assessing qualitatively or quantitatively sorts of labeling, but it is the choice of adequate solution basing on some criteria of selection which are:

- Open tools simple to download on internet, multi-platform and multilingual;
- To permit labeling in French;
- To perform lemmatization and to be equipped with cutter (tokenizer);
- Output files not complex, easy to exploit and very demanded by other extraction tools like YaTea and Gate-Developper.

Basing on these criteria, we have chosen the last French version of Tree Tagger.
The processing of our corpus using Tree Tagger tool generate many files with extension (*.ttg) the content of which is as follows:

Ministère	NOM	ministère
de	PRP	de
la	DET:ART	le
culture	NOM	culture
connaissances	NOM	connaissance
.	PUN	.
Perceptions	NOM	perception
et	KON	et
attitudes	NOM	attitude
de	PRP	de
la	DET:ART	le
population	NOM	population
marocaine	ADJ	marocain
vis-à-vis	ADV	vis-à-vis
de	PRP	de
son	DET:POS	son
patrimoine	NOM	patrimoine
connaissances	NOM	connaissance

Fig. 3 Example of output files, result of processing by the Tree-tagger tool.

Once the corpus is processed, every word is marked without ambiguity and receives one lemma which defines its grammatical category.

2) *Candidate terms extraction*

The extraction of candidate terms, called also operation of terminological extraction, consists of automatic extraction of a list of terms from a specialised corpus. In this regard, there are three categories of tools which depend on the extraction approach used.

- **Tools based on linguistic approach:**

They rely on the location in the text of syntactical scheme of candidate terms. Lexical sequences match the labelled text with pre-set syntactic schemas, for example: syntactic schema [nom nom] /labeled text → (FR: système antipollution | EN: antipollution system), [Nom prep nom] → (FR: tableau de bord | EN: dash board), [nom adjectif] → (FR: champ magnétique | EN: magnetic field).

[4] Cordial: www.synapse-fr.com.
[5] Brill: mail.cst.dk/tools/index.php.
[6] Tree-Tagger: www.ims.uni-stuttgart.de.
[7] Multext: www.lpl.univ-aix.fr/projects/multext/index.html.
[8] Genia Tagger: www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger.
[9] Lia-Tag:.nlp.stanford.edu/software/tagger.shtml.
[10] Pos-tagger: lia.univ-avignon.fr.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. I, Issue 10, December 2013

- **Tools based on statistical approach:**

In this approach, it is the extraction of candidate terms without linguistic analysis of the corpus. The idea is that the more a set of lexical units co-occur, the more this set has chance to form a term. This principle is translated by the notion of mutual information and by the notion of repeated segments.

- **Tools based on mixed approach:**

This type of approach combines two previous approaches. In some tools, obtained results by a linguistic analysis are valid and filtered by a statistical analysis, whereas, in other tools, results of statistical analysis are validated by a linguistic analysis.

In the framework of this paper, we needed to extract terms in French and additioanl results like frequency of occurence of a candidate term and its grammatical composition. The chosen tools should be open sources and use of different algorithm of extraction. For all these reasons, we have chosen two extractors (Yatea and TermoStat) for generation of candidate terms. Then, we have based our methodolgy on examining the exit structures of these two extractors.

- **YaTeA**[11] (Yet Another Term ExtrAtor) (Aubin and Hamon, [19]): is an extractor of terms that identifies and extract nominal groups which may become candidate terms. Every term is analysed syntactically so as to make its structure occur under the form of heads and modifiers. While in the syntactical analysis, an endogenous (from candidate terms of corpus) and exogenous (from external resources) disambiguation is implemented. The extraction of candidate terms is based on hybrid strategy by which the extraction from manual constructed syntactical patterns might be guided and corrected with the help of existing terminological resources. These tested terms help in locating groups and their syntactical analysis. Besides, they participate in the extraction of candidate terms. The necessary linguistic resources to identify and analyse candidate terms are provided for French and English. These resources might be modified by the user. New resources might be also created for a language or a sub-language (technical language of the domain).

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE TERM_EXTRACTION_RESULTS SYSTEM "yatea.dtd">

<TERM_EXTRACTION_RESULTS>
  <LIST_TERM_CANDIDATES>
    <TERM_CANDIDATE MNP_STATUS="1">
      <ID>term7640</ID>
      <FORM>grands centres industriels</FORM>
      <LEMMA>grand centre industriel</LEMMA>
      <MORPHOSYNTACTIC_FEATURES>
        <SYNTACTIC_CATEGORY>ADJ NOM ADJ</SYNTACTIC_CATEGORY>
      </MORPHOSYNTACTIC_FEATURES>
      <HEAD>term439</HEAD>
      <NUMBER_OCCURRENCES>1</NUMBER_OCCURRENCES>
      <LIST_OCCURRENCES>
      <OCCURRENCE>
        <ID>occ2742</ID>
        <MNP>1</MNP>
        <DOC>0</DOC>
        <SENTENCE>727</SENTENCE>
        <START_POSITION>154</START_POSITION>
        <END_POSITION>180</END_POSITION>
        </OCCURRENCE>
      </LIST_OCCURRENCES>
      <TERM_CONFIDENCE>0.5</TERM_CONFIDENCE>
      <TERM_WEIGHTS>
          <WEIGHT name="DDW">0</WEIGHT>
      </TERM_WEIGHTS>
      <LOG_INFORMATION>YaTeA</LOG_INFORMATION>
      <SYNTACTIC_ANALYSIS>
```

Fig. 4 Extract of "candidates file.xml" result of processing by YaTea tool.

The tool YaTea generates a file.xml which contains a list of terms likely to be transformed into concepts. It displays a set of information like number of occurrences and for every occurrence the number of sentence in which the term exists. The offset of the beginning and end of the term permits finding easily terms in the initial text.

TermoStat[12] Drouin, 2003 [20]) is an extractor of multilingual terms (French, English, Spanish and Italian) on line. It is based on linguistic knowledge, and it performs a comparison between use of a term in a specialised corpus and its use in a corpus of general language to determine its pertinence. In the entry, TermoStat selects a textual document. The

[11] YaTea: search.cpan.org/thhamon/Lingua-YaTeA-0.5.
[12] Termostat: www.olst.ling.umontreal.ca/~drouinp/termostat_web.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. I, Issue 10, December 2013

text is grammatically labeled, and morpho-syntactical filters are applied to extract candidate terms. Then, the frequency of every candidate term is compared to its frequency in the corpus of general language. The idea is that the more the gap between the frequency observed in the corpus of analysis and that we can predict from the reference corpus is important, the more the term is potentially interesting. In the exit, TermoStat returns a list of candidate terms giving identified term, its lemmatized form, its frequency and the weight indicating its probable pertinence for the domain of corpus.

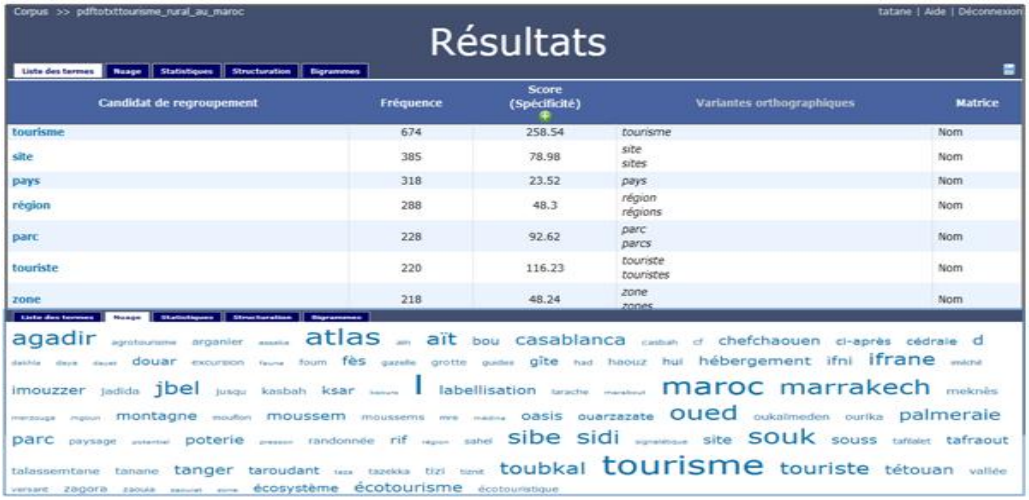

Fig. 5 Example of file result of processing by TermoStat tool.

3) Analysis of Yatea VS Termostat results:

TABLE I
THE EXTRACTION RESULTS OF CANDIDATE TERMS BY YATEA AND TERMOSTAT TOOLS.

	Words/Corpus	**Results Yatea**	**ResultsTermoStat**
Corpus Tourism In Maroc	1 115 922	18662	19488
Rate C.T / M.C		1.672%	1.746%

To measure the performance of these tools, we have calculated the following indicators of quality: (Precision, Reminder and measure function) on the base of created corpus.

With:

- **Precision** = Number of terms correctly extracted / Number of terms extracted
- **Reminder** = Number of terms correctly extracted / Number of terms in the corpus
- **F Measure** = 2 x (Precision x Reminder) / (Precision + Reminder)

TABLE II
CALCULATION OF INDICATORS OF PERFORMANCE FOR YATEA & TERMOSTAT

	YaTea	**Termostat**
Precision	50%	55.1%
Reminder	54.7%	57.9%
F mesure	52.8%	56.5%

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. I, Issue 10, December 2013

After studying these results, we have opted for the extractor of candidate terms TemoStat 3.0.

4) Extraction of conceptual results

From the text, two types of relations might be extracted (Sellami, [21]): **lexical relations** like:

- **Hyperonymy** refers to the relation between a more general and more specific terms. It can be illustrated in the example: "Parmi les mammifères marins, il y a les dauphins" (among sea mammals, there is dolphins). "Dolphin" is then a sort of sea mammals. The general term "sea mammals" is a hyperonymy of the specific term "dolphin".
- **Hyponymy** is the opposite of hyperonymy. For example, "dolphins are sea mammals. In this case, the term "dolphin" is hyponymy of "sea mammals".
- **Meronymy** expresses a part to whole relation between terms. This might be represented by the following example: "Le pain est composé de farine et de levure" (the bread is composed of flour and yeast). In this example, the meronymy between bread and yeast expresses the relation; yeast is an ingredient of bread.
- **Synonymy** expresses a relation of similarity between terms. For example: voiture, auto, bagnole (a car, auto) which are synonyms.

Besides, **transverse relations** refer to specific relations to a domain or general temporary or causality relations. These relations are often expressed with the help of verbs in the text (Grabar et al., [22]).

To extract these relations from text, there are three categories of tools according to adopted approach: tools using syntactic criteria, tools using statistical criteria and others using external resources.

- **Statistical approaches** (Smadja, [23]; Harris, [24]; Bourigault, [25]) are based on the quantity of information in texts and the regularity of expression of sentences in texts. The relations between terms which are identified from the study of contexts related to these terms (words, verbs, adjectives, prepositions, etc.).
- **Syntactic approaches** (Hearst, [26]) supposed to define lexical syntactic patterns. These latters are formed of linguistic units (or markers) indicating the presence of a lexical relation and a set of constraints that the lexical or syntactic context of this marker should fill. For example "comme" (like) is a marker of hyperonymy if it is preceded and followed by nominal adjuncts.

The third category of tools, most recent, suggests the use of **external resources** like WordNet, Wolf, WikiNet, etc. Other works use defined relations in the set ontology for the construction of ongoing ontology.

Researches, notably (Malaisé, [27]), show that methods of relation extraction basing on statistical approaches are not very efficient in the case of corpus with low quantity. However, in our context, we use a corpus with big volume. The statistical methods don't seem to pose problems of use. In the case where the chosen tool does not permit the identification of relation between two terms, we propose the use of internal resources in complementary manner.
In this research domain, there many tools of conceptual relation extraction like Gate-Developper[13], Intex[14], Nooj[15], Unitex[16]. Every tool is based on syntactic, statistical or mixed hypotheses.
Let's consider the following example of lexical syntactic patterns, the aim is identifying and annotating the relations of hyponymy in the text. This pattern is formulated according JAPE syntax ensured by platform of text engineering Gate-Developer.

[13] Gate-developper: www.gate.ac.uk/download/
[14] Intex: http://www.nyu.edu/pages/linguistics/intex/
[15] Nooj: http://www.nooj4nlp.net/pages/download.html
[16] Unitex: http://www-igm.univ-mlv.fr/

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. I, Issue 10, December 2013

Fig. 6 Definition of a lexical syntactic pattern in Gate-Developer for annotating and identifying relations of hyponymy.

When a term is followed by verbal phrase 'C1 is a C2', and itself is followed by other term, then the whole is annotated with semantic relation 'IS_A', and as result a specific processing might be associated.

III. SEMANTIC NORMALIZATION

In this sentence, it is a matter of taking over different parts of the realised model, to verify and complete them. Individual verifications are made on chosen concepts one by one, whereas others, more global ones, tackle roles. These tasks consist of justifying that every element is necessary in the ongoing enrichment of ontology, relevant to this place and according to the purpose of modelling.
Let's state some general principles of normalization before listing different points to be controlled.

- **Unicity of definition:** defining two "identical" concepts in different ways occurs often, because those two terms occur in texts, before noting later that they refer to the same knowledge in the model. In such case, one concept is to be formed.
- **Homogeneity** from points of view of coherence descriptions: for every level of hierarchy of concepts, associated relations to a concept should be chosen in coherence with same point of view. For example belonging relation PART_OF should not be confounded with heritage relation IS_A.

A. VERIFICATION OF THE CONCEPT

For every concept of ontology, answers to these questions might be found whether in texts, by asking experts, or they are unimportant and not been answered deliberately.

1) Validate associated terms, synonymous words of this label.
2) Confirm place of these terms in the hierarchy, by applying differentiation principles (BACHIMONT, [28]).
3) Confirm completeness and unicity from point of view of the principal concept.
4) Verify the whole relations inherited from the root to this concept and to ensure that they define a good level of hierarchy.

B. VERIFICATION OF A ROLE

This verification necessitates listing whole concepts linked by same role, as well as whole roles. Verifying the definition of a relation at the conceptual level refers to:

1) Ensuring that the name of the role will be well interpreted with the given meaning at the moment of its creation and that such meaning is the same as all roles and that the concepts of the domain and their value are well chosen and their cardinality is valid.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. I, Issue 10, December 2013

2) Controlling the role signature, this reverts to pose questions on the possible regrouping of roles, and on their position in the hierarchy.

IV. FORMALIZATION OF THE ONTOLOGY

This last stage consists of translating obtained ontology in the previous stage into a computational specified ontology. In our approach, we have chosen OWL language. This language has advantage of being constituted of three sub-languages and of incremental level of formalization. The use of OWL permits first formalization of ontology and ensures its evolution in time. This language permits also presenting all specified elements according to the needs to which our ontology of the domain is to meet (Moroccan tourism).

```
<?xml version="1.0"?>

<!DOCTYPE rdf:RDF [
 xmlns:TM_VOYAGE_D="&Ontology-khalid;TM_VOYAGE_D&#39;"
 xmlns:TM_SALLE_D="&Ontology-khalid;TM_SALLE_D&#39;"
 xmlns:TM_HONG_KONG="&Ontology-khalid;TM_HONG_KONG,"
 xmlns:TM_SPORT_D="&Ontology-khalid;TM_SPORT_D&#39;"
 xmlns:TM_BATEAU_D="&Ontology-khalid;TM_BATEAU_D&#39;"
 xmlns:TM_UNIVERSITÉ_D="&Ontology-khalid;TM_UNIVERSITÉ_D&#39;"
 xmlns:TM_ÉCHANGE_D="&Ontology-khalid;TM_ÉCHANGE_D&#39;"
 xmlns:Ontology-khalid="http://www.semanticweb.org/Ontology-khalid.owl#"
 xmlns:TM_TOURISME_D="&Ontology-khalid;TM_TOURISME_D&#39;"
 xmlns:TM_PRESQU="&Ontology-khalid;TM_PRESQU&#39;"
 xmlns:TM_SERVICE_DU_TOURISME_A_L="&Ontology-khalid;TM_SERVICE_DU_TOURISME_A_L&#39;">
<owl:Ontology rdf:about="http://www.semanticweb.org/Ontology-khalid.owl"/>

<owl:Class rdf:about="&Ontology-khalid;TM_ABBAYE">
    <rdfs:subClassOf rdf:resource="&Ontology-khalid;TM_MONUMENT"/>
</owl:Class>


<!-- http://www.semanticweb.org/Ontology-khalid.owl#TM_ACCORD_BILATÉRAL -->

<owl:Class rdf:about="&Ontology-khalid;TM_ACCORD_BILATÉRAL">
    <rdfs:subClassOf rdf:resource="&Ontology-khalid;TM_ACCORD_INTERNATIONAL"/>
</owl:Class>
```

Fig. 7 Extract of OWL file of ontology of tourism in Morocco created with Protégé2000 tool.

V. CONCLUSION AND PERSPECTIVES

The progress of web and explosion of stored quantity of information have led to the adoption of formalisms aiming to facilitate the management of knowledge. Ontology is a set of widespread tools that facilitate knowledge sharing. It permits, in fact, to represent concepts as well as relations which link them. However, the pertinence of information it contains depends directly on its updating. This latter consists mainly of adding concepts and relations. It is often realized manually, and makes the task fastidious and the result subjective.
We have used a process which is based on the use tools of languages automatic processing. Since then, the proposed approach permits semi-automatic enrichment of base ontology by adding new concepts and relations.
Results of this experimentation have allowed us to enrich the created base ontology in the framework of GECO-WES project. They are embodied by detection of more than 4412 new key terms of tourism meaning (Concept) and more that 1114 new semantic relations which may link these concepts.

Through our approach we have tried to facilitate the following tasks:

- The creation of textual documents covering studied domain (Tourism in Morocco);
- The pre-processing cleaning of created corpus;
- The linguistic analysis of corpus to identify and collect candidate terms and semantic relations by using appropriate TALN tools;
- Normalization of the ontology and the passage from notion term to concept notion before inserting these concepts in the initial ontology by integrating relations between them;
- Formalization and generation of the domain ontology with protégé2000[17] tool to render it exploitable by other tools of processing and handling of ontology.

[17]protégé2000: http://protege.stanford.edu/.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. I, Issue 10, December 2013

We have noted that handling corpus in the ontological research has an interdisciplinary dimension due to prompt evolution in contexts that modify linguistic needs and uses. Mentioning that this management of corpus called semi-automatic is realized with performing software and subject to human interpretation. That has allowed avoiding the massive use of web which leads to questions of possibilities of controlling texts.
In the rest of this work, we are going to look for solutions to problems related extraction of candidate terms like: ambiguity, anaphora and conference.

REFERENCES

[1] E. Aguirre, O. Ansa, E. Hovy, and D. Martinez , « Enriching very large ontologies using www », in Workshop on Ontology Construction of the European Conference of A.I. (ECAI-00), 2000.

[2] A. Faatz and R. Steinmetz, «Ontology enrichment with texts from the www», in Semantic Web Mining 2nd Workshop at ECML/PKDD, Helsinki, Finland, 2002.

[3] PAREKH V, GWO J.-P. & FININ T, « Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Onto », International Conference of Information and Knowledge Engineering, 2004.

[4] Nédellec C. and Zweigenbaum, " Acts of the day Learning, semantic knowledge and texts ", workshop associated with the platform AFIA ', Laval, in June, 2003.

[5] N Aussenac-Gilles, S Despres, S Szulman, «The terminae method and platform for ontology engineering from texts», Proceedings of the conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, 2008.

[6] Maedche, Kluwer Academic Publishers « Ontology Learning for the Semantic Web», Volume 665 of The Kluwer International Series in Engineering and Computer Science, 2002.

[7] P. Buitelaar, P. Cimiano, and B. Magnini, «Ontology Learning from Text: Methods, Evaluation and Applications», volume 123 of Frontiers in Artificial Intelligence and Applications, 2005.

[8] N Aussenac-Gilles, B Biebow, S Szulman, «Revisiting ontology design: A method based on corpus analysis», 12th International Conference, EKAW 2000 Juan-les-Pins, France, October 2–6, 2000 Proceedings, pp 172-188, 2000.

[9] Philipp Cimiano, Johanna Völker, «Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery », In: Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems, 2005.

[10] S. Mouhim, A. Laoufi, C. Cherkaoui , E.Megder, D.Mammas, "An Ontology based Knowledge Management System For Tourism", Information Systems And Economic Intelligence, Conference SIIE, Marrakech 17-18 février 2011.

[11] Jennifer pearson, « Terms in Context, Studies in Corpus Linguistics », John Benjamins Publishing Company, Amsterdam, 1998.

[12] S. Mouhim, A. El aoufi, M. Eddahibi, H. Eddouzi, C. Cherkaoui, D. Mammass, A Practical and Functional Evaluation of Some Semantic Search Engines, International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 02, No.05, 2012

[13] Kilgarriff, "Comparing corpora", International Journal of Corpus Linguistics 6(1), 97-133, 2001.

[14] Korenius T., Laurikkala J., Järvelin K. & Juhola M., «Stemming and lemmatization in the clustering of Finnish text documents», Conference on Information and Knowledge Management, 2004.

[15] S. Zhang S, O. Bodenreider «Aligning representations of anatomy using lexical and structural methods», AMIA Annu Symp Proc. 2003, 753–757, 2003.

[16] Grefenstette G. , «Exploration in Automatic Thesaurus Discovery», Kluwer Academic Publishers, Londres,1994.

[17] Habert et al., «Integrating Pharmacokinetics Knowledge into a Drug Ontology As an Extension to Support Pharmacogenomics», AMIA Annu Symp Proc. pp 170–174, 2003.

[18] G Pérennou, " The BDLEX project of databases , lexical and phonological knowledge ", first Days of Human-machine GRECO-PRC communication, Publishing EC2, pp 81-111, Paris 1988.

[19] Aubin et Hamon, « Improving term extraction with terminological resources. Advances in Natural Language Processing (Proceedings of the 5th International Conference on NLP (FinTAL'06, LNAI 4139, p. 380–387 : Springer, 2006.

[20] DROUIN, «Term extraction using non-technical corpora as a point ofleverage», citeulike: 1852026, pp. 99-115, 2003.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. I, Issue 10, December 2013

[21] Sellami Zied, Valerie Camps, « DYNAMO-MAS: A Multi-Agent System for Building and Evolving Ontologies from Texts», International Conference on Practical Applications of Agents and Multiagent Systems (PAAMS'12), Salamanca, Springer-Verlag, mars 2012.

[22] Grabar et al., « Combination of endogenous clues for profiling inferred semantic relations: experiments with Gene Ontology», AMIA Annu Symp Proc, 2008.

[23] Smadja, « Retrieving Collocations from Text : Xtract », In Computational Linguistics, n° 19(1), pp 143-178, 1993.

[24] Harris, « Mathematical Structures of Language », Interscience Publisher, Wiley, New York, US, 1968.

[25] Bourigault D., " Upery: a tool of distributional analysis spread(widened) for the construction of ontology from corpus ", Acts of the 9th annual conference on the natural language processing, Nancy, 2002.

[26] M. A. Hearst, « Automatic acquisition of hyponyms from large text corpora», In Proceedings of COLING, Nantes, France, 1992.

[27] Malay V., " linguistic and terminological Methodology for the structuring of differential ontologies from textual corpuses ", PhD thesis, Paris 7 Denis Diderot University, 2005.

[28] Bachimont, B., " semantic and ontological commitment: conception(design) and realization of ontology in engineering of the knowledge. Recent evolution and new challenges ", Eyrolles, Paris, 2000.