An Indexing Approach based on a Hybrid Model of Terminology-extraction using a Filtering by Elimination Terms

Benafia Ali
University of Batna -Batna
Laboratory LIRE -Chaabat Rssas -Constantine
Algeria
Email : ali_bnfa@yahoo.fr

Maamri Ramdane
University of Constantine-Constantine
Laboratory LIRE -Chaabat Rssas -Constantine
Algeria
Email : rmaamri@yahoo.fr

Sahnoun Zaidi
University of Constantine-Constantine
Laboratory LIRE -Chaabat Rssas -Constantine
Algeria
Email : zsahnoun@yahoo.fr

***Abstract*—the extraction of terms is an important step in building a resource for indexing and many powerful tools are available for several languages. This complex process, which identifies candidate terms may become indexes for annotations or documents, is often subject to the problem of lack of relevance of calculated terms. Consequently, the extraction of terminology from the texts must be strong and solid to handle the errors and suggest better results, without encumbering the user with too many proposals of indexes. It is, in this perspective that we propose here a new indexing approach based on a hybrid model of terminologies extraction using a filtering by elimination terms and which operates on a corpus of annotated images with legends .**

***Index Terms*—term extraction, semantic indexing, linguistic analysis, syntactic patterns, complex terms, n-grams .**

I. Introduction

There are two tendencies in the images research models, the research based on the text and the search based on the content. In the textual approach, the image search uses the textual descriptions. Once an image is described by the natural language or by an annotation using keywords, the image search compares the strings of characters corresponding to the image with requests. This mechanism is simple to be implemented, however, in real application, it is often noted that the resources with the keyword are not pertinent and the resources that are pertinent do not contain the same keyword. So, the most obvious problem of the image search based on the text is that it can sometimes reveal too many duplicated images or no image at all. In general, the indexing of images takes little advantage of the many existing work in the textual information search and the natural treatment of languages. The main difficulty faced by the authors is in the selection of the most pertinent indexing terms to describe an image and especially the consideration of the relationships between these terms.

In response to this question, we present in this paper a new paradigm of indexation that combines the language models and the numerical models, which we apply to the case of images provided with legends for the automatic generation of text descriptors. The main purpose of this work is to develop a tool for extracting terms from a corpus of annotated images with legends. These terms issued from the various legends will constitute a linguistic index for the corpus. This index will be added to the descriptors already defined in the model described in [5] to obtain a global index that includes a typology of descriptors. Our contribution is therefore seeking to combine, in a prototype of automatic indexation, terms extracted from texts describing images with visual and semantic descriptors related to images.

We will present in this article firstly the problematic addressed and then situate the perimeter of the study from the context in which we are working. We will describe in the third section an overview of the state of the art of our problem. In the fourth section, we will detail the complete treatment chain which will permit, from the corpus of legends images, to extract a collection of terms serving as

© 2013 ACADEMY PUBLISHER
doi:10.4304/jait.4.1.28-39

a base of linguistic index for the corpus. We analyze in the fifth section the obtained results of our experiments from a corpus and finally we will finish the study with a conclusion and perspectives.

II. Peremeter Of The Study

We find in [5] an approach for indexing images by a semantic graph is based on a representation combining several descriptors. This model like the other models is supported by resources of knowledge and terminology needed for future applications of image processing. The study carried out for the design and the achievement of this model is part of the general framework of the indexation and the search of the images on the Web.

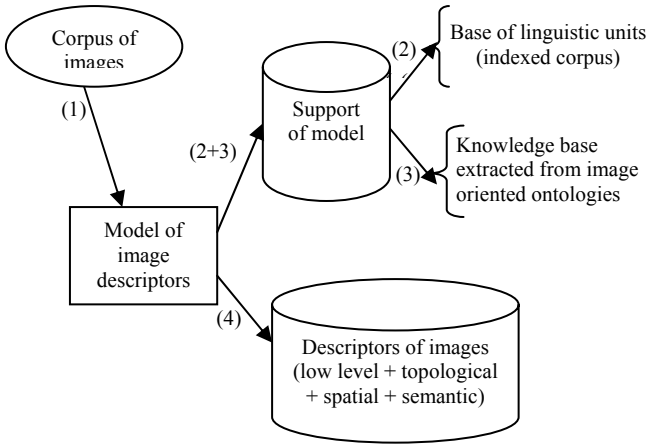

Figure 1. Architecture of global project

The global project of our team aims to develop and validate a methodology of application regarding the indexing and searching of images based on techniques from artificial intelligence, statistical techniques and linguistic techniques. The context of our problem within the scope of this proposal is limited to point 2.

Our problematic is this: we already have a model of description of images [5] where the descriptors associated with the images are all non-textual. Is there a way, and how to expand and consolidate these non-textual descriptors of our model with a structure of textual index?

To respond to this problematic, we will at first segment each legend into sentences, split each sentence recursively into triplets (left syntagm, verb, right syntagm) so that both syntagms obtained become nominal (for each triplet). We extract from these syntagms, the relevant terms which then become descriptors. These multi-terms chosen for the indexation make it possible to designate semantic entities or concepts better than single words, and offer a better representation of the semantic content of a legend.

III. Related Works

Most studies show that indexing by multi-terms offer a distinct advantage because the information searched for is focalised in a privileged way in nominal groups. The different works of extraction of the complex terms from textual corpus use three approaches: statistical or numerical analysis, linguistic or structural analysis and hybrid or mixed analysis. Statistical analysis is based on the study of the contexts of use and the distributions of the terms in the texts; the linguistic analysis exploits linguistic knowledge, such as the morphological or syntactic structures of the terms whereas the hybrid analysis uses the two analyses.

Among the works based on a linguistic approach, we can mention, Lexter [8] who defines a methodology for modelling terminology of the French language. There is also Fastr [16] who uses meta-rules to locate in a corpus, variants of terms from a list of initial terms. [26] exploits the linguistic and lexical features known in the texts in general and apply them to the texts within other areas such as grammatical labelling, extraction of the concepts and the semantic relationships between words. All these methods are based entirely on corpus learning which requires an annotation cost in terms of time and skill required. Even if the linguistic approaches enable obtaining good results, producing electronic dictionaries in which the reuse of semantic information hidden in large databases of terminology, are still too costly in time and effort, to be easily integrated into the part of a flexible approach.

Other studies continue to give priority to the statistical methods, with the introduction of simple language filters to make corrections to noise. There is the Ana system [1] developed as part of work on the automatic indexing of texts in French. The authors describe it as a system dedicated to learning concepts and are therefore not independent of languages and application domains. [12] proposes an interesting method for extracting concepts whose main idea is to measure the correlation in terms of positions of words constituting each concept and the fragments of texts delimited by each concept. We can say that the statistical approaches are therefore faster because they allow to target the subsets of interesting data. They do not require use of external linguistic data to the corpus; they may well do their work in the absence of dictionaries and grammars. This is a definite advantage, because these resources are often the most expensive to develop since they are usually the result of manual work. Despite all these advantages, we can reproach these approaches for the results generated this disability which makes hardly interpretable results in the cadre of a linguistic theory; moreover, they are sensitive to the size of the corpus.

For hybrid approaches, some authors prefer to begin the treatment of the corpus by a linguistic analysis, whose results are filtered using statistical techniques .Others authors proceed to the opposite. [25] describes the Xtract system designed in the domain of automatic indexing and retrieval of the collocations with the detection of complex nominal syntagms. In [7], we find the description of a semi-automatic tool for the identification of complex terms. The proposed approach allows to combine language and digital filters. This approach acquires promising results however it is heavy and off-putting because its interactivity, although the tool proposed can be useful in terminology, indexing and research

© 2013 ACADEMY PUBLISHER

information. These approaches called hybrids also operate the systematicity, rapidity and independence from the domain of statistical algorithms. The intervention of linguistic knowledge and approved statistical calculations allow at these approaches more performance and to obtain more satisfactory results for the linguist on the phenomena which it seeks to observe and describe. This is the reason for which we opted for the choice of this approach.

IV. Methodology

Our methodology aims to select the complex terms from the legends that are associated with images in a corpus. A first selection of these terms is done then filters are applied to this set of terms (composed of noun phrases) to retain at the end of this filtering process that the discriminative terms used for the indexation of the legends. This methodology is then based on a set of filters operating on noun phrases from a set of heuristics, and proceeding by successive elimination of irrelevant terms. The methodology proposed in this paper consists of several steps:

- The first step consists to describe the needs and leads to build a corpus on which heuristics will be developed.

-The second step corresponds to the development of some pre-treatment necessary to annihilate certain ambiguities and inconsistencies in the texts to analyze.

-The third step allows, using analyzers TreeTagger [19] and Syntex [9], to lemmatize the legends, to annotate them and to generate structures of syntactic dependencies.

-The fourth step, is the hinge of the steps considered, present an oriented approach «eliminating by filters" on the extraction of relevant terms.

-The fifth step has as task the pruning of terms and aims to eliminate redundant terms and unnecessary terms resulting from the filtering process. It performs the fusion of all the terms obtained after a series of filters and finally forms the basis of final index.

- Finally, the last step consists to validate our approach based on a set of tests taken from a training corpus.

The methodology we propose aims to enable the development of a text-based index from a set of terms (single words, compound words,) extracted from the corpus of images described by legends. Such a base is more refined and ensures the property of semantic variation. Other stages of our project are not described in this article and are subject to other publication work.

V. Description Of The Steps

A. Preparation of the Corpus

There are two distinct ways to perceive and understand an image. The surface structure is what we see and the deep structure is what the picture really means. Describe the surface structure of the image [15] consists to enumerate the elements that it contains and describe her deep structure , is to explain the meaning of the image and interpret the scenes it contains . Basing on this principle, we consider three types of legends:

- Legends with a simple description and a more precise specification of content of images, done in natural language.

- Legends which describe scenes associated with the images, they express interpretations that we can do for the images and are specified in an uncontrolled free language.

- Legends in the form of keywords combining terms inherent to the image. .

Going from the visual language to the textual language is an extremely difficult task and requires preparation of a long span to be able to represent appropriately the visual content of the image. To annotate an image this is translate what you see in the image, but also what we can highlight from its content , This is where that the process of annotation are really realised. The experts of corpus often put many competing terms to annotate their legends: simple words, compound words, complex words, named entities...

In sum, a good annotation of images from legends must absolutely meet the following requirements:

-distinguish the surface structure from the deep structure of the image during the process of describing.

-take into account the typology of images existing on the Web (photos, cartoons, drawing…).

-choose a vocabulary to identify in an image and describe its objects in a manner more rational

-also, choose a vocabulary and a simple style to locate the type of scene (scene in the countryside, urban landscape, inside, outside, day, night ...).

B. Pretreatments

A term is a unit made up of a word (simple word or compound word) or a group of words. To extract all the relevant terms of the legends, it is necessary to carry out the standardization of the texts of legends.

1) Standardization : it is necessary for a good linguistic and statistical continuation of our treatments. The processing chain corresponding to the normalization is as follows:

Rough sentence → **Treatment of ambiguities/inconsistencies** → Disambiguated sentences → **Recognition of compounds words** → compounds words compounds recognized in the sentence

Figure 2. Processing chain for the normalization of legend sentences

a) Treatment of ambiguities / inconsistencies: we selected for this type of treatment, the cutting of the legends in sentences, the treatment of the capital letters at the beginning of sentences, the elimination of non-compliant symbols (equation, drop cap ...), the elimination of the spaces and the space "- "in the words containing the hyphen (eg. "*porte clefs*" becomes "*porte-clefs*", some treatments of spelling errors (typographical inconsistencies) on a token or more tokens (eg. "*lég ende*" becomes "*légende*" and "*la légende*" becomes "*la légende*") ...

b) Recognition of compound words: the compound words are badly managed with the analyser TreeTagger,

© 2013 ACADEMY PUBLISHER

this is the reason for which we have judge it necessary to build our own processing module of the compound words . In [24], the author shows that it is better to recognize the compound words in texts by consulting the dictionary Delac [23] often very large in size rather than using an algorithm to find the form of a lemmatized compound word from its form arrowed and operative on the morphological categories and inflectional variables of the simple components.

For the identification of compound words, we propose a heuristic which analyzes the terms of the text by group "n-grams" and whenever the term is a compound word we mark it, remove it from the analyzed sentence, we move to following the term "n-grams" and we update the current sentence and so on until there is no further term to deal with. In this method we privilege the short term (composed of 2 or 3 words), this choice is justified by the fact that in the French language, there are more than 70% of compound words that include two or three words, hence the following heuristic:

Algorithme1. Identification of compound words

```
For i from 2 to s do
    While there exist sections to treat do
        While there exist terms of i-grams type
          belonging to the examined section do

          If the examined term is a compound
            word  then Mark_it  Endif

            Skip to the next section
        Endwhile

        Update the section according to the found
                compound words
    Endwhile
Endfor
```

Remark: the *threshold* **s**, corresponding to the allowable score, is a parameter determined empirically by the expert and used for examine the combination of terms to test.

The *marking Mark_**it*** procedure consists in concatenating all the elements that make up the compound word, so that they are in atomic form. With the analyzer TreetTagger , this word is considered as an unknown word. During the second pass of our analyser, it is during the second pass of our analyser that we could then identify the compound word for labelling.

2) *Labeling* : to facilitate the extraction of the terms in the sentence to be examined, you need a morpho-syntactic analysis and a syntactic analysis of surface. Considering the tools available on the web for French and the performances of these, we decided to use TreeTagger [19] for morpho-syntactic annotations and Syntex [9] for annotation of the syntactic dependency relations. The processing chain for this type of analysis is:

Legend standardized → **TreeTagger** → Legend labeled → **Syntex** → Legend labeled and structured

Figure 3. Processing chain for morpho-syntactic labeling

a) Use of TreeTagger[19]: TreeTagger is a system that was developed by H. Schmid in the project "TC" at the University of Stuttgart. It is a tool that allows annotating text with information about the parts of speech (kind of words: nouns, verbs, infinitives and particles) and information of lemmatization. The contribution of TreeTager in our case is to assign to the legend taken from the morphological categories (noun, verb, adjective ...) which facilitates the analysis of the legends linguistically. Our morpho-syntactic labelling procedure including TreeTagger operates according to two passes:

- ✓ *First-pass:* the text to be analyzed in this morpho-syntactic labeling sentence is already standardized moreover the compound words, if they exist, are identified and marked. TreeTagger takes as input a text file that it splits into words by assigning a morphological label to each word and the lemma of the word and then sends back the result of the labeling in the output file.We note that the compound words were been previously treated (cf., recognition of compound words) and are considered for the analyzer TreeTagger as unknown words and the tip used in our analyzer is to concatenate all components of the analyzed compound word in an atomic unit.

- ✓ *Second-pass*: in the first pass TreeTagger achieved labeling of the sentence without treating compound words (considered as atomic words and unknown), it was therefore necessary to make corrections to all these words from the previous treatment of the compound words with the new label NC (lexical label for each component of the compound word) while respecting the output format of TreeTagger.

b) Use of Syntex[9]: .Syntex is a syntactic analyzer of corpus which allows to extract from a corpus a list of names and noun phrases, structured by syntactical dependency relationships. This analyzer takes as input a legend split out in words (simple words and compound words) with each word is associated with a grammatical category (result of TreeTagger). At output, Syntex produces two formats, the first to represent the syntactic dependency relations (subject, direct object ...) between words and the second to represent the networks of syntagms (verbal, nominal, adjectival ...) structured by the relations head and expansion. These two indicators allow us to locate and identify among the examined terms, the substantives, the subjects...

C. Calculation of the Index Terms

1)Modelling in triplet form : the purpose of decoupage of the phrases into noun syntagms is neither used for of comprehension purposes nor for translation purposes, but

© 2013 ACADEMY PUBLISHER

to isolate separately each noun phrase to process its elements. We propose a formalism of representation of the phrases in the form of noun phrases extracted from verbs.

We note D: set of legends, P: set of sentences of the corpus, V: set of verbs that includes the empty set and S: set of syntagms that can be formed from the words.

The description d ∈ D of a legend is formalized as an ordered set of several $p_i \in P$ and $d = [p_1, p_2, p_3 \ldots p_m]$ with $m \geq 1$. A phrase $pi = [t_{i1}, t_{i2}, \ldots, t_{ik}]$ is a set of k noun phrases represented by a triplet $t_{ij} = [S_{JJ}, v_{ij}, s_j + 1\ j]$, with j = 1, k and t_{ij} that can be interlinked or free and $s_{ij} \in S$ and $s_{i+1j} \in S$ the syntagms located to the left and right of the verb $v_{ij} \in V$.

Remark: two syntagms t_1 and t_2 are interlinked if a noun phrase or a part of this phrase that can switch from t_1 to t_2 to form a sentence or a part of the sentence.

2)Procedure of decomposition: as the filters operate on noun syntagms , it is then necessary to transform automatically the textual format of the legends into a canonical format expressed in our model. The method adopted for this kind of formatting consists in analyzing each sentence and to locate from its syntactic structure the verb (the sentence is already labeled by TreeTagger and Syntex). The position of the verb in the sentence allows to split this phrase into two parts, the left part and the right part. We proceed by cutting each sentence around the verb, this decoupage gives two syntagms: a left syntagm and another right and we continue recursively this process of decoupage while the obtained syntagms contain always a verb. The stopping criterion of this recursive process of decoupage into noun syntagms (left and right) stops once the examined syntagm no longer contains the verb.

Legend labelled and structured syntactically →**Decomposition** → Legend in triplet form

Figure 4. Representation of legends in the defined model

Example: we consider the following sentence
« *Cette image contient une grande maison située au bord de la rivière* ».

This sentence is supposed to be already labeled by TreeTagger and Syntex. The decomposition procedure described above turns this specification in form (triplet):

*<cette image, **contient**, une grande maison située au bord d'une rivière>*.

We apply the same process again for the right part of the triplet thus generated; this syntagm is not completely nominal which gives two triplets:
*<cette image, **contient**, une grande maison>*
*<une grande maison, **située**, au bord d'une rivière>*

The process stops because all the syntagms in the left and right of the triplets are nominal. These two triplets obtained are interrelated.

3)Filtering of candidate terms: our approach for the selection of the relevant terms for indexing images from the legends is described by a processing chain using four filters that are presented in the following:

Legend in triplet form →**1st filter** → **2nd filter**→**3rd filter** →**4th filter** → Candidates of retained terms → **Pruning** → Candidates of relevant terms

Figure 5. Processing chain for incremental filtering of terms

• ***1st filter**: structural dependence between terms*

Each term t (n-gram) is written in the form of n words x_1, x_2, x_3, ... x_n with x_i full or empty word. The measures considered here are calculated according to three criteria and allow to express the bond strength existing between the different words xi (full) regardless of empty words, because the frequency of these latter which is much higher than the frequency of the full words. The criteria adopted are:

* criterion of presence: all xi words must be present except the empty words.
* criterion of order: all xi words must appear in the same order xi without taking into consideration of the empty words.
* criterion of distance: all x_i words must be within short distance between them.

-For the first criterion, we propose a new measure of the mutual information MI [10)] for the dependencies between the words x_i treated by pairs (i = 1, n-1):

$$\mathbf{M}(x_i, x_{i+1}) = \begin{cases} \mathbf{freq}(x_i, x_{i+1})/(\mathbf{freq}(x_i)^2 & \text{if } x_{i+1} \text{ is an empty word} \\ \mathbf{freq}(x_i, x_{i+1})/(\mathbf{freq}(x_i)*\mathbf{freq}(x_{i+1})) & \text{if } x_i \text{ and } x_{i+1} \text{ are two full words} \\ \mathbf{0} & \text{if } x_i \text{ and } x_{i+1} \text{ are two empty words} \end{cases}$$

The measure desired for the term T is therefore:

$$\mathbf{IM}(T) = \mathbf{IM}(x_1, x_2) + \mathbf{IM}(x_2, x_3) + \ldots + \ldots + \mathbf{IM}(x_{n-1}, x_n) \quad (1)$$

-For the second criterion, it is to define a weight function on the order of words x_i taken in pairs from term T:

$$\mathbf{P}(m_i, m_j) = \begin{cases} \mathbf{1} & \text{if } m_i \text{ and } m_j \text{ are full and found in that order } (m_i \text{ before } m_j) \\ \mathbf{0} & \text{otherwise} \end{cases}$$

The desired weight is then:

$$\mathbf{P}(T) = \mathbf{P}(m_1, m_2) + \mathbf{P}(m_2, m_3) + \ldots + \mathbf{P}(m_{n-1}, m_n) \quad (2)$$

-Finally for the third criterion, the measure is weighted on the distance between two full words and expresses the bond strength between different words, taken in pairs:

© 2013 ACADEMY PUBLISHER

$$F(m_i,m_j) = \begin{cases} 1/nb(m_i,m_j) \text{ if } m_i \text{ and } m_j \text{ are two full words} \\ \quad //nb : \text{distance in number of full words between the words } m_i \text{ and } m_j // \\ 0 \quad \text{otherwise} \end{cases}$$

The weight for this criterion is:

$$F(T) = F(m_1, m_2) + F(m_2, m_3) \ldots + \ldots F(mi_{n-1}, m_n) \quad (3)$$

Remark: two words situated side by side have a distance equal to 1.

The bond strength of dependency between the words m_i of a term T is:

$$FLD(T) = \alpha_1 * MI(T) + \alpha_2 * P(T) + \alpha_3 * F(T) \quad (4)$$

The coefficients α_1, α_2 and α_3 represent the weightings attributable to criteria defined previously.

Algorithme3. Calculation of the measures of dependence terms

Selected-terms ← ∅
For each sentence p in the legend **do**
For each t_i of p **do**
For j from 2 to card(t_i) **do**
//**card(t_i): number of triplets that make up t_i**//
- Calculate for each term formed by j-grams the score according to the formula (4)
- Select the terms of **s** top scores
//**s: the fixed threshold by the expert** //
- Selected-terms ← Selected-terms ∪ {terms with **s** the top score}
Endfor
Endfor
Endfor

The set of candidates of terms selected by this heuristic are composed by two words (bigrams case), three words (trigrams case) ... They are represented by the set **selected-terms**. This set may contain terms that are included in other words; it must then prune these superfluous terms and retain the terms considered to be most discriminating. This elimination is done according to the scores associated to each term:

Algorithme4. Pruning of superfluous terms

-Sort the elements of the set **Selected-terms** according to the number of words in terms

Final-terms ← ∅ ; trouve←false
While Selected-terms ≠ ∅ **do**
Take a term t_k // **t_k the first term in the list of selected-terms for each passage**//
Search a t_j in the list Selected-terms such that t_k is a sub-sequence of the sequence of t_i

If it exists then **if** score (t_i) ≥score (t_k) **then**
Final-terms←Final-terms ∪{t_i}
Selected-terms←Selected-terms - {t_i}
trouve←true
Endif
Endif
If not trouve **then** Final-terms ← Final-terms ∪ { t_k}
}
Endif
Selected-terms←Selected-terms-{ t_k}
Endwhile

- ***2nd filter:*** *consideration of the builders called " syntactic patterns"*

The previous step has enabled us to make an initial selection of terms formed from 2 words, 3 words by a number of measures. Each candidate thus selected possesses already a lexical labelling (TreeTageer), and to be always kept as a candidate among the selected, it must be submitted to another filter. This second filter is based on the principle of syntactic patterns. We have a base of syntactic patterns which contains a collection of syntactic configurations extracted from an automatic learning of syntactic patterns carried out on a corpus from training. A syntactic pattern is a rule about the order of concatenation of grammatical categories formed for a noun phrase. For our implementation we selected a combination of 128 patterns extracted from a corpus by a manual learning procedure done on a sample of examples of legends. This filter allows to examine each candidate and to check if it is among the existing configurations in the database of patterns built. In the affirmative case, it will be retained otherwise it will be removed.

Example: let us suppose that the candidate elected in the previous step is: <*Au premier plan qui* >.
TreeTagger assigns to the sentence containing this noun syntagm during its analysis the label (Prep, adj, noun, prrel), with the morphological categories: prep (preposition), adj (adjective), name (name) and prrel (relative pronoun).This syntactic pattern does not appear among the set of the existing configurations in the base of patterns, this candidate will then be rejected.

- ***3rd filter****: semantic potential of the syntagms*

To improve and to refine the list of terms obtained during the previous filtering, we consider this time a filter based on a criterion of calculation of measurement concerning the amount of information associated with each candidate term examined. A term with a potential of information sufficient, tends to be a more important and more

© 2013 ACADEMY PUBLISHER

discriminative term in the noun phrase where it is located. This filter assigns a score for each term regardless of the context to which it relates. This score is proportional to the size of terms, their distribution and their type in the corpus, hence:

$$Q_{inf}(t_i) = \sum_{i}^{n} Q(x_i) * FLD(t_i) * \alpha_i \quad (5)$$

with:
-n: number of words that form the term t_i (number of components which form the term t_i).
-Q (x_j): amount of information given to the word x_j.
-Q_{inf} (t_i): cumulative amounts of information for the term t_i.
-FLD (t_i) is a measure of dependence between the different words of the term t_i (cf equation 4).
- α_j: coefficient related to the type of the term t_i.
The values of Q (xi) are those attributed to the grammatical categories of all words found in the corpus. These values are empirically predefined according to their frequency, to their use ....The empty words or uninformative (le, la, de, des ...) used as a tool for the construction of the syntagms have of the measures of quantities considerably lower in comparison to other informative classes.
The coefficients α_j are chosen from the corpus according to the nature of the term and its distribution in the corpus. The value of this coefficient changes according to whether the term is a single word, a compound word, a named entity, a compound term consisting of several words or even composed of associated words. We can find two terms < *système d'information* > and < *système informatique* > such as the word "*système*" does not have the same amount of information depending on whether it is in the 1st or 2nd term because this word (*système*) depends on the distribution of each of these two terms in the corpus.
The named entities, for their part represent a conceptual description which refers to an object whose linguistic representation is often unique; we can describe them as more specific terms, and therefore discriminate. In this case, if a named entity is identified in a noun phrase as a term to be weighted, it will be affected a value sufficient which privilege to be selected as linguistic index term. We can notice from what emerges from this approach of scoring that the longest terms are always preferred. Such a privilege is unquestionably justified by the fact that the longest term remains the most discerning when it is about the choice of descriptors of indexing.
Example: between "*base de données* " and "base de données relationnelles", it is clear that the term chosen is the 2nd term .

Algorithme4. Calculation and selection of discriminative terms

Let L, the list of candidates result of the previous filtering.
-Calculate the score Qinf for each term t in list L
-new-list ← { }

For each term t taken from the list L **do**
If $Q_{inf}(t) \geq s$ then new-list← new-list ∪ { t } finsi
Endfor

• ***4th filter**: full words and empty words*

This type of filter is a heuristic based on the general laws known in the domain of linguistics. In [27], the author confirms that it is possible to build the list of the empty words based on Zipf's law. This law stated in [29] considers that the more the word is frequent, the more it is shorter. The list of the empty words is built on the basis of the lengths and the frequencies of the words. We consider empty words, the words that are both short and frequent. Starting from this principle, we propose for this filter to combine this law to label firstly the analyzed term and to treat the term in question. The labelling for this filter uses two sets of labels: full or informative word and empty or uninformative word. The two criteria used in our non linguistic analyzer are "word length" and "word frequency".
Consider t the term to treat formed by the words m_1, m_2, m_n . The different labels assigned to the words m_i are F (full) and E (empty) and the word is labelled by the label F if it is long and rare and by the label E if it is short and frequent.

Algorithme5. Description of filter

-Labeling each word m_i by examining its characteristic profile (long-short or frequent- rare), this profile is determined according to the frequency of the word in the corpus and its size.
-If the term after labelling has the form $\mathbf{E^+ \alpha_i E^+}$ with $\alpha_i \in$ {F, E} then the words located at the ends are superfluous, we remove the words associated to E^+ from both sides of the term.

To elucidate the task of this filter, we take the example of the following term: <*La grande piscine du village voisin*>.
We suppose that the labeler for this filter gives: E E F E F E

In this example, αi (i=1,3) is represented by the labels F,E,F whereas E+ located on its left is formed by the 2 labels E E and E+ located on its right by the label E.
The words of labeling E E (left) and E (right) will be removed from the chain of term labeling; the candidate term selected is "*piscine du village*".

Similarly, the terms « *de l'image* », « *l'image* », « *une image* », « *à l'image* » are considered identical to the term "image" retained as term reference in the index base.

© 2013 ACADEMY PUBLISHER

D. Pruning of terms

1)Method: : the pruning procedure applies to a collection lists, generated during the filtering phase, which it receives as input. In output, It returns a set of classes of terms constituting the final index..

We note L1, L2, Ln, a collection of n lists containing all the terms selected after a series of filtering. The partitions set of classes of terms which constitutes our index is built progressively and at each iteration we take a term j of the class i (i = 1, n) and we search the nearest class within the meaning in "semantic proximity" to the term j in this set of partitions, if we find him , then we insert the term j in the corresponding class and we update the centroid of conceptual vectors for this class and we repeat the process with the following term in the current list until all the lists are all treated. In the case where the term j does not match any class, we create a new class for this term and we take this class as centroid, the centroid of the term. We note that the construction of this set of partitions of the classes is done without the knowledge of the number of starting classes in other words it's a non-supervised classification of indexing terms. In each class built, we find all the terms having the best measure of semantic proximity with the others

2) General algorithm:

Algorithme6. Research of terms having the best measure semantic proximity

---

- Merge the lists L1, L2 ... Ln in L.

- Eliminate the redundant terms and the terms included in there terms.
- Initialize the index I with the first class from the first term drawn randomly from L

For each term t drawn from L **do**
* Determine the nearest class to the direction of semantic proximity.
* If the required class exists then we insert the term t in this class and we update the centroid of conceptual vectors for this class.
*If the required class does not exist then we create a new class with the term t and we take as centroid , the conceptual vector of the term t.
endfor

---

*3) Explanation**s**:* we clarify in details our algorithm

a) *Fusion of the lists:* this step is necessary because it allows gathering all the terms in order to treat them together and consequently to facilitate the elimination of the duplicated terms.

b) *Elimination of doubled blooms and terms included in other terms*: this operation allows starting from the list previously built to remove all the redundant terms and to remove as well the terms included (non-specific) in other terms. For example between the term « *données* » and « *données numérique*»" we remove «*données* ». The latter problem was already treated during the process of filtering at the local level (sentence). The treatment here concerns the global level (legend).

c) *Initialization: t*he index to be build will contain classes of the most relevant terms. The purpose of these classes is to make our index not fixed in relation to queries on the legends and to solve the problem of semantic variation of the terms. The initialization of the index at the beginning is carried out with the first term (taken randomly from the class **L)** with its conceptual vector.

d) *Determination of the nearest class to the considered term*: we are going to adapt the notion of conceptual vectors [22] for the calculation of measures that allow us to determine the best class to which the term considered will be affected. The notion of conceptual vectors is a formalization of the projection of the linguistic notion of semantic domain in a vector space. It is therefore possible to construct vectors which each component corresponds to a concept. The comparison between two vectors is done using the angular distance. For two vectors x and y to be compared, the angular distance D_a (x, y) is equal to $x*y/\|x\|*\|y\|$ and empirically if this distance is inferior to $\pi / 4$ the thematic proximity is considered near, if this value is higher than $\pi / 4$ the proximity is low and if it is around $\pi / 2$ the proximity is the near zero. From these principles, we will first build for all the terms found in the list L the conceptual vectors and for this purpose we chose an external resource Wolf (Wordnet the Free French) which is a free semantics lexical resource for French. Wolf has been built from the Princeton WordNet and various multilingual resources [18].

To determine the closest class for the term t, we first calculate the angular distance of the conceptual vector of **t** with the conceptual vector centroid of each class and retain the lowest angular distance and as the different vectors have not necessarily the same number of components, we do this calculation on the elements of the neighbourhood thematic [22].

e) *Updating the centroid of the conceptual vectors of the class*: from the centroid of the conceptual vectors of class, we can to project all the vectors of the terms of this class. The centroid of a class contains all the closest terms to the terms belonging to this class. The updating of a centroid consist to take into account the calculated angular distances between the terms with all the elements of the class searched and the centroid of the current angular distances (before inserting the term into the class)… .

We consider, d1, d2, ... dn ,the various angular distances between the terms and all the terms of the class Ci and Ci-1 the centroid of the angular distances of the class before the addition of the new term in this class. The new centroid will have this value :

$$C_i=(\sum_1^i k * C_{i-1} + \sum_1^i d)/(\sum_1^{i+1} k) \qquad (6)$$

The index i represent the *ith iteration* corresponding to the insertion of the term in the searched class.

© 2013 ACADEMY PUBLISHER

Remark: the centroid of the first term to be included in the class (if it is not exists class which does not support this term) is equal 0 (the angular distance of two terms identical is equal to 0).

VI. Experiments And Results

A. *Corpus*

For the experimental protocol, we consider a corpus of 200 images uploaded to the site:
http:/www.pascal-corpus.htm.
We have captioned these images in two distinct ways:

- A first class (**100 images**) contains the short legends (15 to 30 words in average).
- A second class (**100 images**) contains the long legends (more than 100 words).

We excluded in our experimental protocol the analysis of surface annotations legends because this type of description is generally simple and explicit. We chose to test our system with the annotations deeply to better evaluate our approach. Among the recommendations retained on the form and the style granted for the legends, we can list what follows:

- Describe, the best possible, the contents of the image.
- Use the active voice rather than passive.
-Replace the relative subordinate clauses by two sentences.
-Avoid the phrases accumulating numbers, acronyms, abbreviations or enumerations.
-Avoid the clauses, the phrases between brackets, the semicolon or the colon when used as a way to provide accessory information.

The main problem of the evaluation of the indexing, highlighted by [17] is that there is no indexing "reference" perfect as criteria for validate or refute an indexing system, whether it is human or automatic. So, we will use a comparison of the indexing to a "gold standard", a particular index taken as reference, developed by an expert indexer. In this case, the legends of the corpus taken as a test for our indexing approach are subject to the expert to calculate manually all the terms considered pertinent. The validation of the indexing is done by the expert indexer according to its descriptors announced as reference. However, we must remember that indexing is an open problem: for any document, there is not a single and unique set of descriptors constituting an ideal indexing. On the contrary, several solutions are possible and acceptable. The use of a reference index as a basis for evaluation still remains a plausible method to establish an evaluation of the descriptors calculated in relation with those determined empirically. To do this, we collect in what follows the different measurements obtained following three viewpoints of different measures. The use of a reference index as a basis for evaluation still remains a plausible method and often used to establish an evaluated of an indexing system.

B.. Character of Evaluation

To measure the quality of our indexing approach as presented in this article we talk here about an assessment based on indexing compared reference [17]: we have two lists of terms to be compared, one of these lists constitutes the reference with which the other list (output list of our system) must be compared. The model of comparison of the terms of our index with the ground truth index (reference index) uses the following confusion matrix:

TableI.
Matrix confusions for the terms to compare with the terms of reference

evaluated list	∈ reference	∉ reference
selected	Terms ∈ to the list of reference and ∈ to the list to be analyzed	Terms ∉ to the list of reference and ∈ to the list to be analyzed
not selected	Terms ∈ to the list of reference and ∉ to the list to be analyzed	Terms ∉ to the list of reference and ∉ to the list to be analyzed

Let P, R and F_m respective measures of precision, recall and of F-measure, we have:

P= (Terms ∈ to the list of reference and ∈ to the list to be analyzed)/(Terms ∈ to the list of reference and ∈ to the list with analyzer+ Terms ∈ to the list of reference and ∉ to the list to be analyzed) (7)

R= (Terms ∈ to the list of reference and ∈ to the list to be analyzed)/(Terms ∈ to the list of reference and ∈ to the list with analyzer+ Terms ∉ to the list of reference and ∈ to the list to be analyzed) (8)

$\mathbf{F_m}$= (P.R)/(α.R+P (1-α)) with 0≤α≤1, α represents the weight allotted to the precision. (9)

C. Measures and Interpretation

In TableII (see below), the lines represent the various filters to be tested (F1, F1+F2, F1+F3…); the F1 filter is considered as the pivot of this chain of filtering. The method of the n-grams applied to each noun syntagm composed of m words generates m (m-1) /2 terms, and by applying the F1 filter, the terms of weak dependences will be eliminated from the list of the candidates. The remaining filters (F2, F3 and F4) use as input for their treatment, the list returned by F1. The obtained results for this table are not satisfactory because the texts of the legends are supposed to be rough what makes the task of the filters noisy and unfounded since no preliminary analysis (compound words, ambiguities, anaphoras, hooks…) is supposed to be made for this case of evaluation. In terms of short legends (more precise short textual description), the results are definitely better (double yield compared to the long legends) using the vocabulary and the simple style, concrete, direct, and concise. We also notice that the progressive elimination of the candidate's terms by filtering allows eliminating the noise in the selection from the index term.

© 2013 ACADEMY PUBLISHER

TABLEII.
CALCULATION OF THE PRECISIONS, RECALLS AND F-MEASURES FOR A CORPUS OF 100 LEGENDS (SHORT AND LONG) FOR THE SET OF THE FILTERS: F1 (PIVOT OF THE FILTERS), F1+F2 (TEST OF THE 2ND FILTER)… F1+… F4 (ALL FILTERS) WITH NOISY TEXTS

	Long legends				Short legends		
	P (%)	R (%)	F_m (%)		P (%)	R (%)	F_m (%)
F1	40.09	42.66	41.33		24.14	26.82	25.40
F1+ F2	52.19	40.37	45.52		31.28	32.21	32.26
F1+F 3	49.78	49.94	49.85		25.67	29.81	27.58
F1+F 4	51.91	52.50	52.20		28.43	26.63	27.50
F1+F 2 +F3+ F4	57.62	58.24	57.92		35.89	38.74	37.26

TABLEIII contains results on the indexing obtained by our approach by assuming this time that the legends have been standardized and labelled manually by the expert. The texts submitted to our indexer are supposed treated and containing no errors. The results for the extraction of indexing terms are being improved compared to the previous case. We obtain for this type of evaluation how to estimate the true character of the filters regarding their efficiency and their performance under the ideal conditions that is to say, when the corpus is pretreated in advance by the expert. The success rate for the case of short legends is quite acceptable since it reached the limit of 80% though if we consider the filters one by one, their rate is around 73% which proves that their contribution is very interesting. The weakness of our approach always remains the indexation of legends developed since the threshold reached remains more or less enough (around 57%) whereas filters taken separately give a 52%. Considering the results obtained in this case, we can say that our approach of identification of terms for indexing improves the quality of results. Noise filtering is then dependent upon a manual pretreatment. This is why the competence of an expert is also important.

TABLEIII
CALCULATION OF THE PRECISIONS, RECALLS AND F-MEASURES FOR A CORPUS OF 100 LEGENDS (SHORT AND LONG) FOR THE SET OF THE FILTERS: F1 (PIVOT OF THE FILTERS), F1+F2 (TEST OF THE 2ND FILTER)… F1+… F4 (ALL FILTERS) WITH MANUAL PRETREATMENT OF THE TEXTS

	Long legends				Short legends		
	P(%)	R(%)	F_m(%)		P(%)	R(%)	F_m(%)
F 1	63.58	62.18	62.87		41.64	42.39	42.01
F1 +F 2	74.97	72.57	73.75		53.17	53.97	53.56
F 1+ F3	75.32	71.09	73.14		52.32	54.15	53.21
F 1+ F4	72.50	73.45	72.97		52.88	51.01	52.01
F1+F2+ F3+F4	77.60	78.63	78.11		57.29	56.47	56.87

Finally, TABLEIV contains the results obtained from real tests of our approach since the previous treatments of standardization and labelling preceding the filters were also considered. We note that if we take all the filters together, we reach an acceptable threshold of 70% for short legends and a threshold of 54% for the long legends especially in the case of a fully automated approach for a chain of treatment composed of a phase of standardization, labelling and various filters considered. The results obtained in this case are less reliable compared to previous results (automatic procedure VS manual procedure) that are why we plan to integrate to our implementation a learning system, operating on the basis of corpus for the phase of pretreatment.

TABLE IV.
CALCULATION OF THE PRECISIONS, RECALLS AND F-MEASURES FOR A CORPUS OF 100 LEGENDS (SHORT AND LONG) FOR THE SET OF THE FILTERS: F1 (PIVOT OF THE FILTERS), F1+F2 (TEST OF THE 2ND FILTER)… F1+… F4 (ALL FILTERS) WITH AUTOMATIC PRETREATMENT OF THE TEXTS

	Long legends				Short legends		
	P (%)	R (%)	F_m (%)		P (%)	R (%)	F_m (%)
F1	53.38	52.54	52.95		32.27	34.80	33.17
F1+ F2	65.76	64.23	64.98		42.71	43.34	43.02
F1+F3	64.50	63.01	63.74		41.58	46.22	44.00
F1+F4	63.84	61.47	62.20		40.96	43.39	42.63
F1+F2 +F3+ F4	69.53	70.20	69.86		53.17	54.73	53.93

D.. Consideration of the Pruning Phase of the Terms

The previous results do not hold the step of terms pruning. The processing which constitutes the chain of analysis of legends operates sentence by sentence and extracts the candidate's terms for the indexing. This list of terms can be redundant, including the ones with the others… Pruning intervenes only for the set of the four filters. The fusion of all the lists resulting from the sequence of filters for each sentence (belonging to the same legend) and their pruning slightly improves the results in terms of high precision (cf. TABLEV).

© 2013 ACADEMY PUBLISHER

TABLEV.
CALCULATION OF PRECISIONS, RECALL AND F-MEASURES FOR A CORPUS OF 100 LEGENDS (SHORT AND LONG) FOR ALL FILTERS (TAKEN INCREMENTALLY) AFTER PRUNING OF SELECTED TERMS

	Short legends				Long legends		
	P(%)	R(%)	F_m(%)		P(%)	R(%)	F_m(%)
F1+ F2+ F3+ F4	70.99	71.28	71.13		53.84	56.20	54.25

VII. CONCLUSION AND OUTLOOK

We proposed in this article a new paradigm for indexing based on the notion of syntagmatic term. This paradigm involves the translation of nominal phrases extracted from the corpus to be indexed into syntagmatic index terms having a complete semantic. The method used is hybrid and it combines numerical and linguistic filters. These filters are not expensive in time of calculation and easy to implement. This is an induction to improve the accuracy rates. The results of the filtering process still remain well relative, they depend on the size of the corpus (more the corpus is bigger, more the results are better) and depend on the pre-treatments done on the texts (recognition compound words, elimination of ambiguities ...) and depend especially on the nature of the legends. In other words, a short text with a style of description more simple may lead to better results and accuracy more efficient.

We have also solved the problem of semantic variation of the set of terms filtered by an unsupervised algorithm of classification, based on a structure of conceptual vectors. On the other hand, with the envisaged prospects for the further of this work, several aspects should be mentioned:

-Addition of a treatment of pruning of the terms from the corpus (global analysis) by merging the different legends between them

- Adding a treatment concerning the recognition of the named entities since the named entities are considered more easily identified in the texts and can be chosen as indexing terms.

-Consideration of the unknown words.

-Integration of a treatment on the identification of collocations.

- For the long legends, it is strongly recommended to define a heuristic on the thematic salience that will enable to approach the other treatments without ambiguity.

-Construction of a referential of terminology for access and the search for information in particular the images through their legends.

-Definition of a learning system for the construction of the conceptual vectors for our corpus even though the problem still remains difficult

REFERENCES

[1] Anguehard, C.(1993), Acquisition de terminologie à partir de gros corpus, Informatique & Langue Naturelle, ILN'93, Nantes, p.373-384,

[2] Aussenac N.G , Jacques M.P(2008) , Designing and Evaluating Patterns for Relation Acquisition from Texts with CAMÉLÉON. Dans : Terminology, John Benjamins Publishing Company, Amsterdam, Numéro spécial Pattern-Based approaches to Semantic Relations, Vol. 14 N. 1, p. 45-73.

[3] Baziz M., Boughanem M., Aussenac-Gilles N.(2005), Conceptual Indexing Based on Document Content Representation , Ed., Informa tion Context : Nature, Impact, and Role : 5th International Conference on Conceptions of Library and Information Sciences, CoLIS, vol. 3507, Lecture Notes in Computer Science.

[4] Benafia A., Maamri R., Sahnoun Z., Construction of a terminological ressource support based on ontologies for images descriptions model , to appear in proceeding,the International Conference on Informatics & Applications , Malaysia, June 3-5, 12. .

[5] Benafia A., Maamri R., Sahnoun Z , Saadaoui S. ,Saadna Y.(2011),A Representation model of images based on graphs and automatic Instantiation of its Skeletal Configuration ,Springer-Verlag Berlin Heidelberg 2011 ,Intelligent Interactive Multimedia Systems and Services Smart Innovation, Systems and Technologies, Volume 11.

[6] Benoît S., Darja F. (2008), Construction d'un wordnet libre du français à partir de ressources multilingues, TALN 2008, Avignon,

[7] Biskri I., Meunier J., Joyal S.(2004) , L'extraction des termes complexes : une approche modulaire semi-automatique JADT 2004 : 7es Journées internationales d'Analyse statistique des Données Textuelles ,

[8] Bourigault D.(1994), Lexter , un logiciel d'entrainement de terminologie , Application à l'extraction des connaissances à partir des textes. Thèse en Mathématiques, Informatique appliquée aux sciences de l'homme EHESS, Paris.

[9] Bourigault D.(2007), un analyseur syntaxique opérationnel :SYNTEX Université Toulouse-Le Mirail Mémoire présenté pour l'obtention d'une Habilitation à Diriger les Recherches

[10] Church K , Hanks P. (1996) Word association norms mutual information and lexicography computational p 22-29.

[11] Courtois B.(2011), Dictionnaire electronique du LADL pour les mots simples du Français , DELAS V06/2 , Université Paris VII .

[12] Dinh, D. and Tamine, L.(2011), Biomedical concept extraction based on combining the content-based and word order similarities. In SAC, pages 1159–1163.

[13] [Enguehard, C. (1992). ANA, Apprentissage Naturel Automatique d'un Réseau Sémantique. Thèse de doctorat, Université de Technologie de Compiègne.

[14] Evans D.A., Ginther-Webster K. ,Hart M. ,Lefferts R.G., Monarch I.A.(1991), Automatic indexing using NLP and first order thesauri, RIAO'91 Recherche d'Informations Assistée par Ordinateur .

[15] Grimault F.(2009), Terres d'Innovation Photographie, Indexer et légender, Recueil INRA.

© 2013 ACADEMY PUBLISHER

[16] Jacquemin C.(1991) , Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus, Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes.
[17] Lancaster, F. W.(1991) Indexing and abstracting in theory and practice , University of Illinois : Champaign, Hardcover, 328 pages.
[18] Sagot B., Fiser D.,Building(2008), a Free French WordNet from Multilingual Resources. Proceedings of the 11th international conference on Text, Speech and Dialogue Springer-Verlag Berlin, Heidelberg .
[19] Schmid H.(1995), Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.
[20] Schwab D.(2001), Vecteurs Conceptuels et Fonctions Lexicales : application à l'antonymie. Mémoire de Dea, Université Montpellier II - Sciences et techniques du Languedoc.
[21] Schwab D., Lafourcade M, Prince V.(2002) ,Amélioration de la représentation sémantique lexicale par les vecteurs conceptuels: le rôle de l'antonymie: 6es Journées internationales d'Analyse statistique des Données Textuelles.
[22] Schwab D., Lian Tze, L., Lafourcade M.(2007), Conceptual Vectors, A Complementary Tool To Lexical Networks. In: Proceedings Of The 4th International Workshop On Natural Language Processing And Cognitive Science (NLPCS 2007), Funchal, Madeira , Portugal .
[23] Siberztein M.(1990), Le dictionnaire électronique des mots composés en langue Française , pp. 71-83, Paris : Larousse.
[24] Siberztein M.(1999), Traitement des expressions figées avec Syntex In analyse lexicale et syntaxique : le système Intex pp 425-449 Publishing Compagny : Amsterdam/Philadelphia .
[25] Smadja F.(1993)Retrieving collocations from text: Xtract. Journal Computational Linguistics, Vol. 19, N°1, p. 143-177.
[26] Stevenson M.,Guo Y.,Gaizauskas R.& Martinez D.(2008), Knowledge sources for word sense disambiguation of biomedical text. In BioNLP, p. 80_87, Columbus, Ohio, USA.
[27] Vergne J.(2005) , Une méthode indépendante des langues pour indexer les documents de l'internet par extraction de termes de structure contrôlée. In Actes de la Conférence Internationale sur le Document Électronique (CIDE 8), Beyrouth, Liban .
[28] Zamin, N. and T.B., Baharudin, B (2009) Machine Learning Algorithms for Text-Documents Classification" **Journal of Advances in Information Technology** (**JAIT**) Volume : 1 Issue : 1 February 2010, ISSN : 1798-2340 'The Development of Phrase-Based Transfer Rules"
[29] Zipf G.K. (1968), The Psycho-biology of Language. An Introduction to Dynamic Philology. The M.I.T. Press, Cambridge

© 2013 ACADEMY PUBLISHER