Text Concepts Extraction based on Arabic WordNet and Formal Concept Analysis

Nadia Bouhriz
University Hassan II
Faculty of Sciences Ben M'sik
Department of mathematics and computer, BP.7955, Sidi Othmane, Casablanca, Morocco.

Faouzia Benabbou
University Hassan II
Faculty of Sciences Ben M'sik
Department of mathematics and computer, BP.7955, Sidi Othmane, Casablanca, Morocco.

Habib Benlahmer
University Hassan II
Faculty of Sciences Ben M'sik
Department of mathematics and computer, BP.7955, Sidi Othmane, Casablanca, Morocco.

ABSTRACT
In the context of Information Retrieval System (IRS), semantic coherence between text and the terms chosen to represent them, provides precision in the answers returned to the user. So, for the improvement of the capacity of these systems, it is necessary to design and develop methods based on a semantic text processing, for choosing the appropriate terms, which can represent semantically the contents of this text. We will be interested by this area of research in this paper, more particularly, we propose a method, allowing the extraction of concepts which represent the semantic content of an Arabic text. These concepts are extracted, from Arabic WordNet (AWN), which we apply for their, afterward, Formal concept Analysis, to produce a set of concepts, more reduced and more relevant.

Keywords
Terms Extraction, Concepts Extraction, Arabic WordNet, Formal Concept Analysis, Arabic Text.

1. INTRODUCTION
The indexing process consists in extracting terms, from texts and queries, which we call "index terms" or "descriptors", they are used afterward for selecting, the relevant texts to a specific query, according to the commons index terms between them. The classic approach adopted by IRS [1] for this process, is based on keywords(without senses). These systems are consequently, generator of a reduction in, precision and recall, on their answers. this is due, in particular, to the polysemy and synonymy, who are present in the natural language.

To deal with these problems, the researchers adopted a new approach which consists on the integration of semantic aspect during the indexing process. It is what we call semantic indexing or conceptual indexing, so the elements of the index, in this case, can be terms of a thesaurus, concepts of an external semantic resource, summaries of paragraphs, or any other unity which can express semantics contained in the text. Our work is focalize in this area of research. More particularly, we propose a method that allow the extraction of concepts which are going to represent the semantic contents of an Arabic text, these concepts are extracted from the lexical resource AWN, according to the terms present in the text, and afterward a process of disambiguation is applied for the ambiguous terms (having many senses) for identifying exactly their appropriate sense, and finally, we apply, for these concepts, Formal Concept Analysis that aim increasing the capacity of our approach to produce a set of concepts; more reduced and more relevant, and especially given a best representation of the semantic text content.

This paper is organized as follows: In section 2 we introduce some basic definitions on Formal Concept Analysis. In the section 3, we present the Arabic lexical resource AWN. The section 4 is devoted to some related work. In section 5, we describe the proposed approach. In Section 6, we show the algorithm of the proposed approach. And finally, we have finished in section 7 , with a conclusion and our future work.

2. FORMAL CONCEPT ANALYSIS
The Formal Concept Analysis (FCA) is a mathematical formalism, based on the order theory, and applied on the analysis of data [2]. More particularly, the FCA is a process which allows to discover all the possible groupings of objects having common properties. The central notion of the FCA is the formal context. It is a triplet $\mathrm{K} = (\mathrm{G}, \mathrm{M}, \mathrm{I})$, Where G is a set of objects, M a set of properties and I a binary relation on $\mathrm{G} \times \mathrm{M}$ expressing that an object have a property.

From a Formal context, we calculates the formal concept where, a formal concept is a pair (E, I), where E is the maximal set of objects (called extension) possessing all the properties of I, and I is the maximal set of properties (called intension) shared by all the objects of E.

The formal concept $\mathrm{C_K}$ of the context $\mathrm{K} = (\mathrm{G}, \mathrm{M}, \mathrm{I})$, is partially ordered by the inclusion of the extensions. This relation called, relation of specialization between concepts, is noted $\preccurlyeq_{\mathrm{K}}$, formally she is defines as follows: $\mathrm{C_1} = (\mathrm{A_1}, \mathrm{B_1}) \preccurlyeq \mathrm{C_2} = (\mathrm{A_2}, \mathrm{B_2}) \leftrightarrow \mathrm{A_1} \subseteq \mathrm{A_2}$ ou $(\mathrm{B_2} \subseteq \mathrm{B_1})$, we say that $\mathrm{C_1}$ is more specific than $\mathrm{C_2}$.

The set of all concepts with the partial relation of specialization $\mathrm{L} = (\mathrm{C_K}, \preccurlyeq_{\mathrm{K}})$ is a complete lattice, called concepts lattice or Galois lattice. In our work, we are going to call it Galois Lattice to avoid confusion with the AWN concepts.

The following table, gives an example of a formal context: G is a set of six objects $(O_1, O_2, O_3, O_4, O_5, O_6)$, and M is a set of seven properties $(P_1, P_2, P_3, P_4, P_5, P_6, P_7)$:

Table 1. Example of a Formal Context

$G \times M$	P_1	P_2	P_3	P_4	P_5	P_6	P_7
O_1	×	×	×	×	×		
O_2	×	×	×				
O_3	×	×	×				
O_4	×			×			
O_5		×	×			×	×
O_6		×	×				

The lattice L can be represented by a Hasse diagram in which the nodes are the concepts, and edges are the links of specialization/generalization.

The following Figure illustrates the Galois lattice corresponding to the formal context given previously:

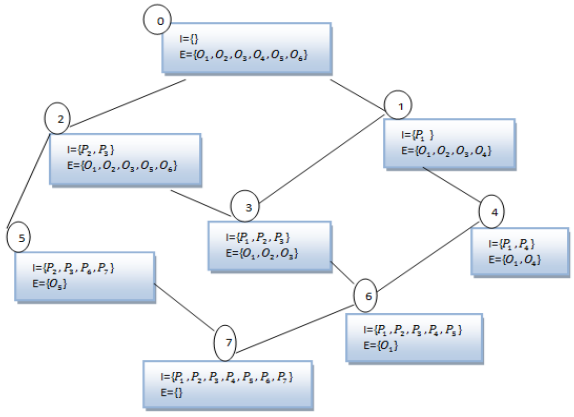

Figure 1. Galois Lattice associate to Table 1

The general concept contains all objects; his intension is empty because no property is common to all objects. also, the specific concept is defined by all properties; in our example, her extension is empty, because no object is described by all properties.

3. ARABIC WORDNET

Arabic WordNet (AWN) is a lexical resource for standard modern Arabic based on WordNet of Princeton used for English. AWN is based on the conception of Princeton WordNet universally accepted (PWN). It is a linguistic resource with a formal semantic foundation which is able to support the richness of Arabic language as described in [5].

The AWN have a structure of thesaurus, he is organized around the structure of synsets, which is a set of synonyms and link describing relation with other synsets. Every word can belong to one or several synsets, and to one or several categories of the speech. The AWN is so a lexical network where the nodes are the synsets and relations between synsets are the edges.

Finally, AWN is a resource for the general Arabic language available online. It currently has 11269 synsets and 23,481 words. [3], [4], [5], [6], [7], [8].

4. RELATED WORK

The use of lexical resources (taxonomy, ontology, dictionary ....) to solve problems of semantic variations contained in a text, was the subject of several works, which allowed them to remove the constraints of synonymy and polysemy. On the other hand, the use of these lexical resources is often accompanied by a disambiguation phase for affecting the appropriate sense to an ambiguous word. the most precise approaches, that have been proposed for this aim, is that based on the calculation of the semantic distance: there are two types for measuring this distance; the first, based on the edges (edge-based) [9] and the second based on the nodes (node-based) [10]. in our work, we will using, for calculating this distance, the first one, more particularly, we are going to consider the shortest path between two nodes, equal to the number of the existing edges in this path [9].

In the case of Arabic, two recent studies, are adopted the approach using lexical resources:

- Mohammad Wedyan and al. 11] were based on the technique of reformulating queries to find similar terms to the initial query, they based, for that, on a thesaurus which they automatically built from a set of documents. This experience has show an improvement in the performance of the system from 10 to 20%.
- Mohammed Alaeddine Abderrahim and al. [12], were used the AWN to extract concepts, who will represent documents and queries during the indexing process in an information retrieval system for the Arabic language, to improve the quality of the search. They experiments proves an improvement in precision and recall in the responses of the system.

At the same time, several studies [13], [14], [15], were used the FCA to construct a formal context from text, for different aims. In the case of Arabic, [16] were used the FCA to transform a text into a structured data space, the elements of this space are concepts extracted from a financial ontology in Arabic. this approach has show good results in time execution and number of trivial concepts explored.

In this paper our contribution is to improve the relevance of the terms chosen to represent the meaning contained in an Arabic text. the novelty of our proposed approach is, firstly; the localized manner in which the processing is done, as described in the next section, secondly, the use of FCA aims to filter the final set of concepts that will represent the text by removing the concepts that have minor importance in the in the meaning of the text and keeping those that have major importance.

5. PROPOSED APPROACH

Our proposed approach aim, extracting concepts, representing an Arabic text, this concepts can be used later as semantic descriptors for indexing phase. This approach, is based , firstly, on AWN, and secondly, on the FCA formalism. it's split into three steps :

- **First Step** : Terms Extraction (Figure 2)

This step consists in extracting terms representing the text. This will be done as the following way: we will segment each text in paragraphs, afterwards, these paragraphs will be segmented in sentences, then we apply for each sentence a process of removing empty words, we obtain a list of no stop words $W = (W_1, W_2, W_3, \dots, W_n)$, after that, each word in this list will be tested by verifying is it belong to a synset in AWN, or not : if it is, the word will be added to the list of terms $T = (T_1, T_2, T_3, \dots, T_n)$, if it's not, we add her stemming form to the list T. For the Stemming step we are based on the work of [17], which provides a method for Stemming Arabic words, based on finite state automata. Finally, we obtain the list T, which is the set of terms, representing the sentence, ordered according to her position in this sentence.

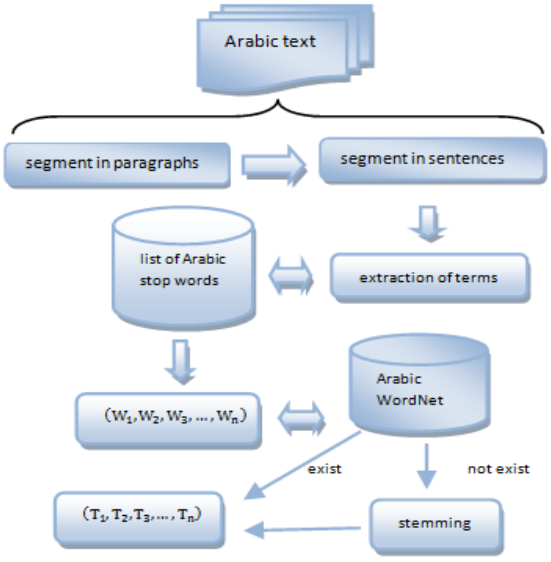

Figure 2. Terms Extraction Step

- **Second Step**: Concepts Extraction (Figure 3)

In this step, we project terms in the list produced during the first step, on AWN, in order to return synsets that they belongs. So, we proceed as follows: for each term, in the list T, projected on AWN; three cases are possible: term belongs to a one synset (non ambiguous terms), or to more than one synsets (ambiguous term), or else to any synset. Terms that do not belong to any synset will be kept as they are and will be added to the list of concepts extracted $C = (C_1, C_2, C_3, \dots, C_n)$. With regard to terms that belongs to one synset, we will extract the name of the concept and we add it to the list of extracted concepts C. Finally, the terms that belongs to more than one synset, are disambiguated, by using the semantic distance, to choose the appropriate synset that contain the exact sense of this term. Particularly, we will proceed as follows: for each ambiguous term (belong to many synsets), we calculate the semantic distance for each of these synsets with the synset to which belongs, his nearest non ambiguous term in the sentence; the synset who have the smallest semantic distance will be the synset which contain the appropriate sense for the ambiguous term, the concept associated to this synset will be add to the list of extracted concepts. Note that if, in a sentence, there is no non ambiguous terms, then in this case we look in the nearest sentence.

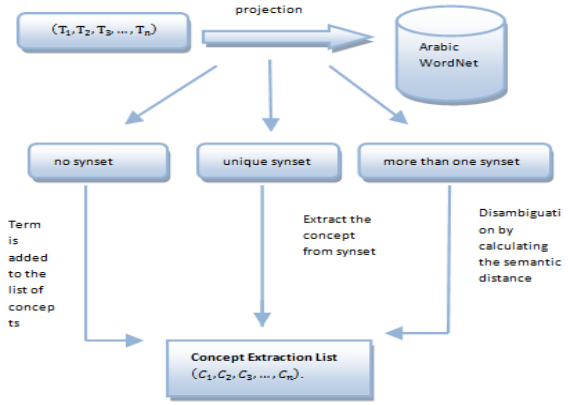

Figure 3. Concepts Extraction Step

Example. for this set of terms ["شخص","توظيف","ساهم","سياحة","قطاع"], our approach return the associate synsets for each terms as follows :

{[قطاع], [قطاع, مقطع]}

{[سياحة]}

{[أعطى, ساهم, تبرع], [أضاف, منح, أضفى, ساهم], [قدم, تبرع, تحمل, زود, ساهم, دعم, أعان]}

{[استثمار, توظيف], [تعيين, توظيف, تنصيب, تولية, تسمية, تخصيص], [استخدام, استعمال, توظيف], [استخدام, توظيف]}

{[شخص, فرد], [نفس, شخص, ما, مخلوق, شخص, روح, فرد, أحد ما, إنسان], [امرؤ, مرء, شخص, إنسان]}

the unique terms who's non ambiguous is سياحة, after applying the disambiguation phase, we obtain for each ambiguous term his appropriate synset as follows :

[قطاع]

[أضاف, منح, أضفى, ساهم]

[تخصيص, تسمية, تولية, تنصيب, تعيين, توظيف]

[شخص, فرد]

- **Third step:** Generate Final Concepts List (Figure 4)

Our contribution is specifically focused in this step, we propose the use of FCA to increase the ability to detect all final concepts that have a significant degree in the semantic representation of the text, and this, by building a conceptual representation of the binary relation concepts × paragraphs. particularly, we will start from concepts extracted in the second step, we build a formal context, extracting for each concept it represents paragraph in the text. Objects in this formal context are then the concepts and properties are, the set $(\in P_1, \in P_2, \in P_3, \dots, \in P_q)$, which mean, belonging to the paragraphs of the text . The idea is to choose, after the construction of Galois lattice, the formal concept which his intension contains the largest number of properties; i.e. representing of the largest number of paragraphs of text. Its extension will contain the set of concepts that we choose for representing the semantic content of the text.

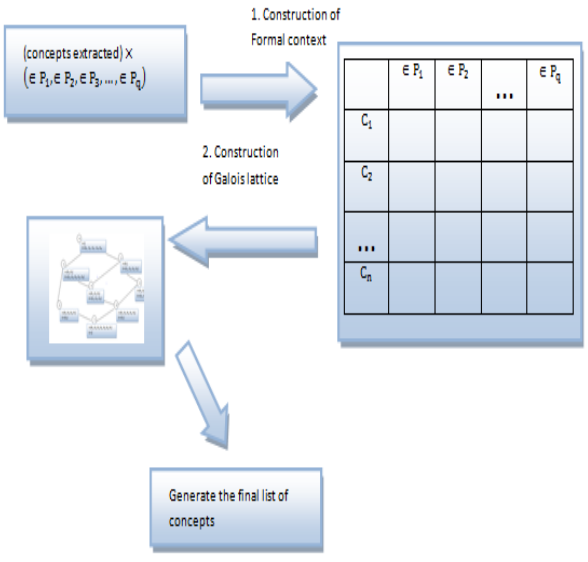

Figure 4. Generate Final Concepts List

6. THE ALGORITHM OF THE APPROACH

Input : Arabic document D

Output : set of document concepts(CD)

$P=\{P_1, P_2,..., P_n\} \leftarrow$ ParseOnparagraph(D)

For i=0 to n **do**

$S_i=\{s_{i1}, s_{i2},.., s_{im}\} \leftarrow$ parseOnsentences(p_i)

For j=1 to m **do**

$T_{ij}=\{t_{ij1}, t_{ij2},..., t_{ijq}\} \leftarrow$ ExtractTerm(S_{ij})

End For

For j=1 to m **do**

$\mathbf{C_{ij}=\{C_{ij1}, C_{ij2},..., C_{ijq}\}} \leftarrow$ ExtractConcpt(T_{ij})

End for

$C_i=\bigcup_j C_{ij}$

End for

ConstructLattice(D)$\leftarrow$ FormalContxt($\bigcup_{i=1}^{n} C_i, \bigcup_{i=1}^{n} \in i$)

$C_D \leftarrow$GeneratListConpt(ConstructLattice(D))

ParseOnparagraph(): function which parse an Arabic document in a paragraph

parseOnsentences() : function which parse a paragraph in sentences

ExtractTerm() : function which extract terms of a sentence

ExtractConcpt(): function which extract concepts corresponding to a set of terms.

ConstructLattice() : construct the Galois Lattice of a document

FormalContxt(): construct the formal context of a set of concepts and set of paragraphs of a document.

6.1 Algorithm of Term Extraction

Input : S_{ij}, the j^{th} sentence of the i^{th} paragraph

Output: Set of Extracted terms

W$\leftarrow$ parseOnWord(S_{ij})

for k = 0 to W.length **do**

while(W(k)!= stop word)

{**if** (W(k)$\in$ AWN)**then**

$T_{ij} \leftarrow$ W(k)

else

$T_{ij} \leftarrow$ Stem(W(k))

End if

}

End for

Algorithm of Concept Extraction

Input : Set of Term T_{ij}, of the j^{th} sentence of the i^{th} paragraph

Output : Set of the extracted concepts

Step of Classification of terms

TA_{ij} :set of ambiguous terms of sentence S_{ij}

TNA_{ij} :set of non ambiguous terms of sentence S_{ij}

$TNCA_{ij}$:set of terms who's not belonging the AWN of sentence S_{ij}

CA : function which returns number of concepts associated to a term

for each t $\in T_{ij}$ **do**

if (CA(t)=1) **then**

$TNA_{ij} \leftarrow$ t

else if (CA(t)>1) **then**

$TA_{ij} \leftarrow$ t

else

$TNCA_{ij} \leftarrow$ t

End if

End for

Step of disambiguation

`TA_ij_disam :`Set of the appropriate concept of ambiguous terms of sentence S_{ij}

`disamb() :`Function who's disambiguise the ambiguous term by calculating the semantic distance with the ambiguous term and the near non ambiguous term in the sentence S_{ij}

`Near_TNA():`Function who returns the near non ambiguous term for a term t

while($TA_{ij} \neq \emptyset$)

if ($TNA_{ij} \neq \emptyset$**)then**

for each t $\in TA_{ij}$

TA_{ij}_disam $\leftarrow$ disamb(Near_TNA(TNA_{ij},t),t)

end for

else

while(TA_{ij}_disam $\neq \emptyset$**)**

if ($TNA_{ik} \neq \emptyset$ && k $\neq$ j) **then**

for each t $\in TA_{ij}$

TA_{ij}_disam $\leftarrow$ disamb(Near_TNA(TNA_{ik},t),t)

end for

end if

end while

end if

end while

Extraction of concepts from AWN:

`extract_concpt():` function which extract the appropriate concept of a term

while ($TNA_{ij} \neq \emptyset$)

for each t $\in TNA_{ij}$ **do**

$C_{ij} \leftarrow$ extract_concpt(t)

end for

end while

while ($TA_{ij} \neq \emptyset$)

for each $t \in TA_{ij}$_disam **do**

$C_{ij} \leftarrow$ extract_concpt(t)

end for

end while

while ($TNCA_{ij} \neq \emptyset$)

for each $t \in TNCA_{ij}$ **do**

$C_{ij} \leftarrow t$

end for

end while

7. CONCLUSION AND PERSPECTIVES

Based on studies done in the context of extracting concepts based on AWN, we can say that his contribution is remarkable in diverse areas, including the IR where the adoption of these approaches characterized by a reduction of silence and noise in IRS answers.

In this paper, our main objective is to combine this approach which exploits the AWN lexical resource with the mathematical formalism of FCA in order to increase the ability to extract a set of concepts, reduces and relevant, describing exactly the semantic content of an Arabic text. the extended objective of our work is to use this set of concepts as a semantic descriptors for this texts, in order, to improve the performance of an IRS for the Arabic language.

A lot of testing and improvements of the proposed approach are still needed, particularly on the properties of the formal context (concepts × paragraphs): other properties, which can enhance the representativeness of a concept in the text, must be looked for. secondly, the lexical resource used for the extraction of concepts is very important in the effectiveness of our approach, which requires to look for other resources and to test them. As perspective, an evaluation of the proposed approach, for a corpus of Arabic documents, in terms of efficiency of the concepts extraction task is the subject of our next paper.

8. REFERENCES

[1] Ricardo A. Baeza-Yates, Berthier A. Ribeiro-Neto: Modern Information Retrieval ACM Press / Addison-Wesley 1999.

[2] B. Ganter and R, Wille, Formal Concept Analysis. Springer- Verlag, 1999.

[3] Horacio Rodríguez, Sabri Elkateb, William Black, Piek Vossen, Adam Pease, Christiane Fellbaum: Building a WordNet for Arabic, http://www.adampease.org/Articulate/publications/LREC.pdf , 2006.

[4] Musa Alkhalifa : Arabic WordNet and Arabic NLP. JETALA 5-7 June, Rabat 2006.

[5] Christiane Fellbaum, William Black, Sabri Elkateb, Antonia Marti, Adam Pease, Horacio Rodriguez, Piek Vossen : Constructing Arabic WordNet in Parallel with an Ontology.

[6] http://www.globalwordnet.org/AWN/meetings/meet20050901/Fellbaum.ppt 2005.

[7] Sabri Elkateb, William Black, Piek Vossen, David Farwell, Adam Pease, Christiane Fellbaum.: Arabic WordNet and the Challenges of Arabic. http://www.mt-archive.info/BCS-2006-Elkateb.pdf (2006).

[8] William J. Black, Sabri ElKateb: A Prototype English-Arabic Dictionary Based on WordNet. http://www.fi.muni.cz/gwc2004/proc/95.pdf , 2004

[9] William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, Christiane Fellbaum: Introducing the Arabic WordNet Project http://www.globalwordnet.org/AWN/meetings/GWApaper.pdf , 2006.

[10] Pedersen T., Patwardhan S. and Michelizzi J. : WordNet::Similarity - Measuring the Relatedness of Concepts. Proc. of HLT-NAACL, pages. 38-41, 2004

[11] Resnick P. : Using information content to evaluate semantic similarity in a taxonomy. Proc. of the 14th International Joint Conference on Artificial Intelligence, pp. 448–453, 1995

[12] Mohammad Wedyan , Basim Alhadidi and Adnan Alrabea; The effect of using a thesaurus in Arabic information retrieval system IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 1, November 2012

[13] Mohammed Alaeddine Abderrahim, Mohammed El Amine Abderrahim, Mohammed Amine Chikh ;using Arabic WordNet for semantic indexing in information retrieval ;JCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 2, January 2013

[14] Rokia Bendaoud, Mohamed Rouane Hacene, Yannick Toussaint, Bertrand Delecroix et Amédéo Napoli : Construction d'une ontologie à partir d'un corpus de textes avec l'ACF. *Dans* Frankie Trichet, éditeur : Actes des 18èmes Journées francophones d'Ingénierie des Connaissances (IC2007). Cépaduès, 2007.

[15] Philipp Cimiano, Andreas Hotho et Steffen Staab : Learning concept hierarchies from text corpora using formal concept analysis. Journal of Artificial Intelligence Research (JAIR), pages 305-399, 2005

[16] Donald Hindle : Noun classification from predicate-argument structures. *Dans* Proceedings of the Association for Computational Linguistics, pages 268–275, 1990.

[17] F. Ferjani, S.Elloumi, A.Jaoua, S.Ben Yhaya, S.A.Ravan, J.Jaam, N.Semmar : feature extraction based on isolated labels application for automatic news categorizatio. Proceeding of the CITALA'12, pages 91-99, 2012.

[18] Y. El younoussi A. Sdigui Doukkali, E. Ben Lahmar, "La racinisation de la langue arabe par les automates à états finis (AEF)" in " International Colloquium on Arabic Language Processing" CITALA2007 18-19-Juin 2007.