International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.1 (Nov-2021)

http://dx.doi.org/10.12785/ijcds/1001114

AR2Concept Automatic Extraction Concepts From Arabic Text Language

Zineb Kheira Bousmaha Ossoukine[1], Hafsa Oulhaci [1] and Lamia Hadrich Belguith[2]

[1] Department Computer Science, FSEA, Laboratory RIIR, University Ahmed Benbella, Oran 1, Algeria
[2]Department Computer Science, FSEGS, Laboratory MIRACL, University Sfax, Tunisia

Received 6 February. 2020, Revised 21 March. 2021, Accepted 3 April. 2021, Published 25 Nov. 2021

Abstract: Our objective is the design and realization of an automatic system of extraction of concepts from a text in the Arabic language as a first step towards the creation of ontology.

The architecture we adopt is an original approach for Arabic language texts that combines the semantic concept extraction method based on the Latent Semantic Analysis documentary search technique with the K-means algorithm.

Faced with the problem posed by the K-means algorithm for the number of clusters to be fixed, we propose a solution that we have evaluated on a set of texts. The first results are satisfactory.

Our AR2Concept system allowed the identification of concepts with an f-measure rate of around 80%.

Keywords: ANLP, Ontology, Concepts Extraction, LSA, K-means

1. INTRODUCTION

In recent years there has been a huge increase in the amount of Arabic electronic information, and access to relevant information has become increasingly complex. The realization of a concept extraction system seems interesting and constitutes an area in its own right which lies at the intersection of the automatic processing of the Arabic language (ANLP), the search for information (RI) and the creation of ontology. Researches in this area already exists for other languages such as English, French but for Arabic, it's just beginning. A question arises: What methods should be used or adapted to extract automatically candidate terms from Arabic text?

Various methods of extraction concepts exist in the state of the art, we count the methods based on external resources [1], methods of "frequent pattern mining" [2], methods based on WordNet and centrality [3] and the Latent Semantic Analysis (LSA) method [4]. Existing theories do not provide as rigorous a simulation as the LSA that imitate the human categorization of words and human judgments. It offers a cognitive model of the representation of the meaning of words, and it estimates the comprehension and coherence of texts. It represents a challenge that has been solved for other languages unfortunately no work has been done in this sense for the Arabic language [5].

The comparative study by [6] discusses the related work on ontology learning, specifically the extraction of concepts and relationships from Arabic textual resources. It clearly shows the use of linguistic approaches such as the use of syntax rules and the use of external semantic resources (such as ontologies thesauri, AWN dictionary ...), statistical approaches based on sequence frequencies and hybrid approaches for extracting concepts. We notice that no work has used the LSA for the extraction of concepts despite its performance for other languages.

We propose an original approach of extraction of concepts for Arabic texts. We combine the techniques of ANLP to extract knowledge from texts and we use the LSA method that we hybridized with the K-means algorithm. This combination was necessary for the conceptualization phase in cluster recovery. Faced with the problem posed by K-means for the number of clusters to be fixed, we proposed a solution that we evaluated on several texts. The first results are satisfactory

E-mail address: kzbousmaha@univ-oran1.dz , hafsa7oua@gmail.com , l.belguith@fsegs.rnu.tn

The rest of this paper is organized as follows: an overview of recent existing works in concept extraction is summarized in section 2. Section 3 describes our approach, explaining related methods such as LSA and K-means clustering. An experimental study showing the effectiveness of our approach is presented in Section 4. Finally, section 5 concludes the paper and discuss areas of future work

2. RELATED WORK

The most recent decade has witnessed an increasing concern for building automatically Arabic ontologies from text. Several works already exist in this area for different languages such as English and French [7]. Unfortunately, for the Arabic language, works are still in the beginning [8].

To build an ontology, the first step is to find the important concepts of the domain. Concept extraction from Arabic documents has been a challenging research area, because, as opposed to term extraction, concept extraction is more domain-related and more selective [9]. In the literature, we find several studies used for term extraction which can be a single word or multi-word term. However, we note that only the studies of [10], [14] and [15] extract concepts.

[10] Present a method for concept extraction from free Arabic text documents. The method applies a set of rules, which are determined by an external agent, on a raw Arabic text. The process for the method consists of two main phases namely: extracting patterns that extracts the features of the training concepts to generate two sets prefix and postfix, and testing phase which discover a set of learned concepts using the extracted features obtained from the previous phase. In the experiments, they used 50 agricultural documents containing a description regarding field crops. The results are around 66% f-score

[11] Analyzed the text with AraMorph, they extracted simple terms with TF-IDF measure and applied a rule-based approach and a statistical method of mutual information (MI) for extracting collocations. They reached a precision value of 86% from Quranic Corpus.

[12] Used the linguistic filter to extract multi-word term (MWT) candidate matching given syntactic patterns to extracted Arabic terminology. The TF-IDF is applied to rank the single word terms candidate and statistical measures (PMI, Kappa, Chi-square, T-test, Piatersky-Shapiro and Rank Aggregation) for ranking the MWTs candidates. The authors indicated a precision value of 80% from the Islamic corpus.

[13] Considered contextual information and both termhood and unithood for association measures at the statistical filtering. To extract MWT candidates, they applied syntactic patterns and they used C-value, NC-value, NTC-value and NLC-value for candidates ranking. The NLC-value measure outperformed others with precision value 82% on an environment Arabic corpus.

[14] Describe an approach starting first with the collection and preparation of the corpus through normalization, removing stop words and stemming; then, to extract simple and complex terms they use a statistical method called “the repeated segment” then, they apply “a weighting filter” to select terms with sufficient weight. The weight is measured by the weighting method term TF–IDF to remove some terms that are not considered as concepts of the domain. The results show a precision equal to 89% on average, and a recall equal to 82% on average, from three resources: a corpus of Arabic texts and two external resources: an Arabic dictionary of synonyms and antonyms and the lexical database Arabic WordNet.

[15] Uses a combination of linguistic, statistical and domain knowledge to extract domain-relevant concepts from Arabic texts, to determine the relevance of these candidates within the domain, different statistical measures are implemented (CF-DF, TF-IDF, CF-IDF, RCF, Avg-CF). They demonstrate that Concept Frequency-Document Frequency (CF-DF) algorithm is the best choice for concept extraction and proposed a new one for candidates weighting. Domain knowledge is integrated into the module, it is obtained from a domain-specific list extracted from the taxonomy of the corpus. The concepts will be displayed to the expert to choose the valid concept and to add the missing one. In order to evaluate the performance of the method, they focus on the medicine domain from hadith corpus. According to the authors, the results were satisfactory.

[16] Affirm that concept extraction can be broken into the identification of all possible concepts and selection of the most important ones. Previous methods start from the idea that concepts can be found as a word using grammatical or syntactical information, then a more sophisticated version of the statistical idea is the use of LSA. LSA is a technique in natural language processing or the analysis of the topics in a document. It is argued that a document consists of several topics, some of which are described intensively in several sentences, and hence form the major content of the document

[17] LSA's purpose is to “distill” from the corpus a set of relevant “concepts”. Each concept is described by three characteristics: the relevance of the concept for each document in the corpus, the affinity for the terms present in the corpus and the usefulness of the concept in

describing the variance of the data. By selecting only the most important concepts, LSA can describe the data with an approximate representation that uses fewer concepts, which eliminates "noise" and merges similar subjects [4]. By using LSA, concepts in the document can be identified in singular vectors, one sentence per vector was then retrieved.

Different approaches using LSA for task-related with ontologies can be found in literature and the results are very satisfactory [16] [18] [19]. A detailed state of the art on this researches is given [18].

LSA is a featurization step. After featurization, we can use a clustering algorithm. Clustering is important in data analysis. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). The only thing we need for a clustering algorithm is a distance metric. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. In order to use K-means, we need a representation of the items you want to cluster and a similarity measure between two items representations. Given these, K-means will be used successfully [20].

Our approach is based on this idea and it was adapted and appliqued to extract the ontological concept automatically from the Arabic Text.

3. CONTRIBUTION

The global architecture of AR2Concept is represented by fig.1. In order to lead to the generation of concepts from a text in the Arabic language, a series of processes is executed.

In preparing for LSA, the domain corpus is divided into sentences and the empty words are removed. The performant morphological and disambiguation analyzer online MADAMIRA [21] is applied. An annotated XML file is generated. This file contains a set of sentences, where each sentence is composed of a set of words preprocessed by this parser. At this level, we have made some modifications concerning the terms of the annotated XML file compared to the word, the lemma and the form when the word is not recognized. Then we extract the simple terms and compound terms from the annotated XML file. This analysis is utilized to reduce the dimension of LSA. In the next step, we proceed at the application of the LSA algorithm. The output of this algorithm is semantic vectors with k dimension for each word identified. We apply a global weighting function to each to improve retrieval performance. We classified the results of LSA by using the K-means clustering method then we extract the pertinent concept.

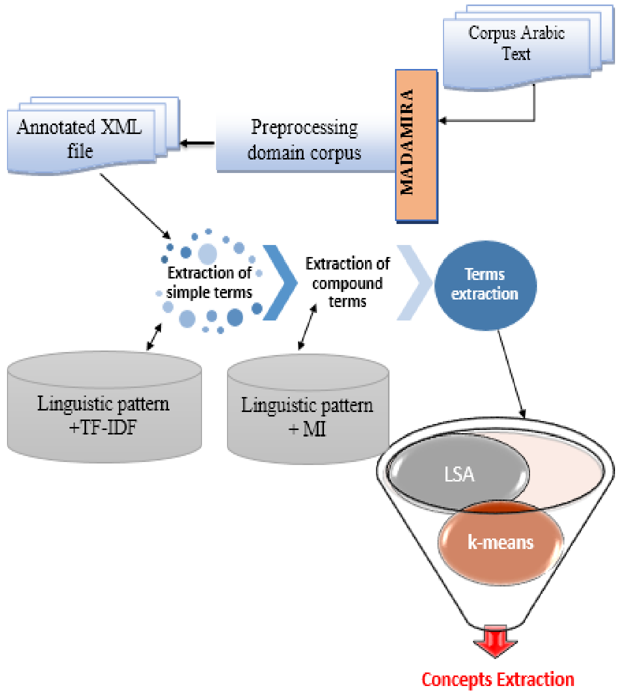

Figure 1. AR2Concept Architecture

A. *Extraction of simple terms*

We begin by grouping words into lemma, then by eliminating empty words. For the extraction of the simple terms, we recover only the words of the text whose form is NN, DT + NN, NNS, DT + NNS, NNP, DT + NNP where the part of speech is noun, noun_prop which means names and proper names. Note that NN represents the noun. NN with P in the plural and S in the singular forms, JJ an adjective and DT the determinant.

We proceed to the weighting of the extracted simple terms. A term adequately represents the context only if its degree of importance in this context is significant. In order to weight these simple terms we use the TF-IDF measure. We apply TF-IDF weighting because it removes the effect of high frequency terms on the importance of a context.

B. *Extraction of compound terms*

We combined a linguistic approach based on syntactic patterns and statistical filtering based on mutual information (MI). We are interested, in our design, by collocations formed of two lexical units and respecting the schemes: (NN) + (JJ); (NN) + (DT + JJ); (NN) + (DT + NN); (NN) + (NNP); (NN) + (DT + NNP); (DT + NN) + (DT + JJ). Example: أوراق حميدة, جسم المصاب, خلايا الجلد

For each pair of previously obtained terms, we calculate the MI. Given two terms designated by the variables x and y. Their probabilities of observation are respectively p (x) and p(y). The probability of observing them together is p (x, y). The frequencies of x and y are respectively f (x), f (y), and the MI is written in the following form (1):

$$MI (x, y) = p (x, y) / (p (x)*p (y)) \quad (1)$$

$$\text{With } p (x) = f (x) / N$$

$$p (y) = f (y) / N$$

$$p (x, y) = (f (x, y)/ N$$

Where: N is the total number of words in the text

C. *Extraction of concepts*

Concepts are generally only a set of terms. Terms are words or sequences of words (compound terms) that can be retained as entries in an ontology. In order to build the core of the ontology, we followed the three steps proposed by our approach as shown schematically in fig.2.

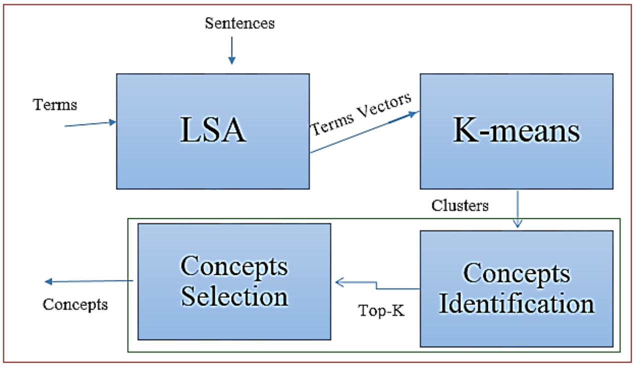

Figure 2. Approach to concept extraction

1) 1st step: Latent Semantic Analysis

LSA is based on the following definition: two contexts "Sentences" are similar if they contain similar terms, and two terms are semantically close (similar) if they appear in similar contexts. Each word or term is represented by a vector in a multidimensional space of very large size by means of a statistical analysis. All the words of the corpus are thus represented by a vector and the similarity between two words corresponds to the cosine between the two vectors of these words. Similarities can also be calculated between groups of words, or between a word and a group of words.

The phases of our analysis with the LSA are:

a) Transformation: LSA takes as input a voluminous texts, and builds afterwards a matrix of occurrences W represented after having done the pretreatment (removal of the words stop, lemmatization ...). In our matrix n * m, the lines represent the n terms and the columns the m contexts (sentence) with n >> m. w_ij: the number of occurrence of the term i in the context j

b) Matrix weighting: this step consists in normalizing the occurrence matrix using TF * IDF or other weighting (log-entropy, BM25, ...). we use the formula (2), (3),(4).

c) Singular Value Decomposition (SVD): allows the cration of semantic links. LSA applies a singular value decomposition of the matrix W. This matrix W decomposes into a product of three matrices: $W = USV^T$; the matrix U represents the terms, the matrix V the documents (sentences in our conception) and the matrix S the singular values SVD that yields a simple strategy to obtain an optimal approximation for a using smaller matrices. If the singular values in S are ordered descending by size, the first k largest may be kept and the remaining smaller ones set to zero.

d) Reduction of matrices to k dimensions: The initial matrix is reduced by nullify the coefficients of the diagonal matrix S, starting with the smallest. A compression of the information is thus obtained by the selection of k dimensions among r. Only the first k singular values (the largest ones) are preserved. The number of dimensions has been determined empirically by different tests, an inappropriate number can leads to a loss of information, or does not make it possible to emerge the semantic links between the words [18]. This number is an open issue in the latent semantic analysis literature. In our research, we took retaining only the first two ranks by keeping the first two columns of U, V, and S. In the case of large text corpora, k could be augmented [18].

e) Comparison: Once the final matrix has been obtained, the semantic similarities between words or groups of words can easily be calculated. This vector representation makes it possible to represent any sequence of words by a vector sum of the vectors of the words that compose them. Each word or group of words is defined by a vector, and to calculate the proximity between these two vectors there are several methods [22]. The best known measure is the cosine of their angle.

Except the last step (comparison), we apply the four phases of the LSA of the transformation until the reduction of matrix to k dimensions.

Table 1 shows a simplified example of sentences and Table 2 its transformation into a matrix of occurrences.

TABLE I. EXAMPLE SENTENCES IN ARABIC

Label	**Sentences**
Sent1	ذهب أحمد إلى الجامعة Ahmed went to the university
Sent2	خرج المدرس من فصل The teacher left a class
Sent3	ذهب الرجل إلى البيت The man went home

TABLE II. OCCURRENCES MATRIX

Terms	Sent1	Sent2	Sent3
جامِعَة/ university	1	0	0
مُدَرِّس/ teacher	0	1	0
فَصْل/ class	0	1	0
رَجُل/ man	0	0	1
بَيْت/ home	0	0	1

Regarding the second step, we propose to standardize the weighting as follows in (2), (3), and (4):

$$TFIDF_{ij} = TF_{ij} * IDF_i \quad (2)$$

$$\text{With } TFij = 0.7 + 0.7 * (TFij / \max TFij) \quad (3)$$

$$IDFi = \frac{\log \frac{N+0.5}{n_i}}{\log(N+1)} \quad (4)$$

TF_{ij} is the frequency of the term i in the document j with double normalization 0.7 and n_i is the number of contexts (sentence, paragraph) in which appears the term i; N is the total number of contexts

Note that we arrived at this formula after studying several formulas whose details in section 4. Applying this standard TF-IDF transformation to our example we obtain the matrix represented in Table 3.

TABLE III. OCCURRENCE MATRIX NORMALIZED BY THE STANDARD TF-IDF

Terms	Sent1	Sent2	Sent3
جامِعَة/ university	0.9036	0.6325	0.6325
مُدَرِّس/ teacher	0.6325	0.9036	0.6325
فَصْل/ class	0.6325	0.9036	0.6325
رَجُل/ man	0.6325	0.6325	0.9036
بَيْت/ home	0.6325	0.6325	0.9036

The third step is the key to the method. The singular value decomposition of the standardized occurrence matrix W of our example is a product of three matrix: W = USV^T

Where the matrix U corresponds to the terms, the matrix V to the sentences and the matrix S contains the singular values.

$$U = \begin{pmatrix} 0,443822 & -5,55112e-17 & -0,896115 \\ 0,448058 & -0,500000 & 0,221911 \\ 0,448058 & -0,500000 & 0,221911 \\ 0,448058 & 0,500000 & 0,221911 \\ 0,448058 & 0,500000 & 0,221911 \end{pmatrix}$$

$$S = \begin{pmatrix} 2,80142 & 0,00000 & 0,00000 \\ 0,00000 & 0,383393 & 0,00000 \\ 0,00000 & 0,00000 & 0,296788 \end{pmatrix}$$

$$V = \begin{pmatrix} 0,547802 & 0,00000 & -0,836608 \\ 0,591571 & -0,707107 & 0,387354 \\ 0,591571 & 0,707107 & 0,387354 \end{pmatrix}$$

And finally, by applying a dimension reduction with k = 2 to our example, we obtain the following result:

$$U = \begin{pmatrix} 0,443822 & -5,55112e-17 \\ 0,448058 & -0,500000 \\ 0,448058 & -0,500000 \\ 0,448058 & 0,500000 \\ 0,448058 & 0,500000 \end{pmatrix}$$

$$S = \begin{pmatrix} 2,80142 & 0,00000 \\ 0,00000 & 0,383393 \end{pmatrix}$$

$$V = \begin{pmatrix} 0,547802 & 0,00000 & -0,836608 \\ 0,591571 & -0,707107 & 0,387354 \end{pmatrix}$$

We applied the LSA method up to phase 4 in order to find the correlations between terms and associate them with concepts. Then, a problem appeared: how to recover the set of concepts from the results obtained. We, therefore, thought that the solution would be the classification of term vectors. Hence the application of the K-means algorithm. Regarding the following treatment with the application of the K-means algorithm, we keep only the term vectors taken from the matrix U.

2) *Second step: Algorithm of K-means (K-means)*

The K-means algorithm is a data partitioning algorithm for statistics and machine learning (specifically unsupervised learning).

The phases of the algorithm are:

1. Initializations for k group centers (centroids).

2. Find for each data point, the nearest class center using the Euclidean distance criterion.

3. Replace each centroid with the average of data points in its group.

Repeat 1+2 until convergence, that is until the centroids of the classes or groups do not change.

Identifying the number of clusters or classes often remains an open problem. But can be solved according to other parameters such as the maximum diameter of classes or try different k values and choose the k that optimizes a criterion of quality / validity of the clustering obtained empirically. But whatever the solution, it was always necessary to interrupt the process to introduce the value of k or to set it from the beginning, which is not obvious. So we propose a novel solution.

During our researches on the LSA method, we noted a formula that calculates the value of k relative to the rank of the matrix for size reduction (fourth step of LSA) [23]. We were inspired by the idea by using in our calculations the exploiting of data (the information) concerning the preceding steps (extraction of the simple and compound terms). Our formula (BO19) was applied to all the texts of the corpus and the results were correct compared with manually entered k.

BO19 formula (5) is:

$$k = \frac{\sqrt{NB\_MP - NB\_TS - NB\_TC}}{2} \quad (5)$$

Where:

- NB_MP: the total number of words and punctuation marks (".", ",", ":", ...) in a text
- NB_TS: the total number of simple terms
- NB_TC: the total number of compound terms

The number 2: represents the two previous steps namely the extraction of the simple terms and the extraction of the compound terms.

The Table 4 shows the 2 clusters obtained automatically without any interruption, after the application of K-means algorithm for our example:

TABLE IV. CLUSTERS OBTAINED AFTER APPLICATION OF K-MEANS ALGORITHM

Cluster1	Cluster
جَامِعَة	مُدَرِسٌ
رَجُلٌ	فَصلٌ
بَيْت	

3)Step 3: Identifying Concepts/selection Concepts

Once this classification of Clustered Terms is complete, the system must extract the most relevant terms to form concepts. These concepts will serve as a basis for the creation of the core of the ontology. It extracts the concepts by calculating the centrality of a term in its cluster, that is to say the term whose average distance with all the other terms of the cluster is minimal. In this case, we obtain a concept per cluster while terms recognized as concepts were not taken. To remedy this, we sorted the average distances calculated previously in ascending order and we recovered the TOP-k. We have chosen a value of k for value 3. This value has been set after several tests.

The Table 5 shows the TOP-k concepts obtained by cluster of our example:

TABLE V. TOP-K CONCEPTS OBTAINED BY CLUSTER

Concept 1	Concept 2
رَجُلٌ	مُدَرِسٌ
بِنتٌ	فَصلٌ
جَامِعَة	

4. EXPERIMENTATION AND DISCUSSION

AR2Concept is developed in Java using the NetBeans environment. To build the corpus, we used texts available on Internet from Altibbi portal http://www.altibbi.com/. It is an online Arabic medical and health resource. The documents provides under أخبار ومقالات طبية/ Medical news and articles.

Our experiment was carried out on 5386 texts whose average of the words by sentences is of 25. At final a list of concepts are proposed automatically after entering text.

At this level, we applied our mathematical formula, in order to recover the exact number of clusters to generate. The concepts are extracted and selected.

We calculated the precision and the recall by applying the formulas (6) and (7):

$$\text{Accuracy} = VP \, / \, (VP + FP) \qquad (6)$$

$$\text{Recall} = VP \, / \, (VP + FN) \qquad (7)$$

Where

VP: true and taken by the system, FP: false and taken by the system, VN: true and not taken by the system and FN: false and not taken by the system

We have selected nine formulas for the weighting of the occurrence matrix (see Table 6) on these texts in order to extract the concepts in such a way that the most informative terms are privileged in order to have a good evaluation of the relevance of a term in a text of the corpus.

TABLE VI. FORMULAS FOR THE WEIGHTING OF THE OCCURRENCE MATRIX

Formulas term weighting	
1	$TFIDF_{ij} = TF_{ij} * IDF_i$ with $TFij = 0.7 + 0.7 * (TFij \, / \max TFij)$

	$IDF = \frac{\log \frac{N+0.5}{n_i}}{\log(N+1)}$
2	$TFIDF_{ij} = TF_{ij} * IDF_i$ with $TFij = 0.8 + 0.8 * (TFij / \max TFij)$ $IDFi = \log \frac{N}{n_i}$
3	$w_{t,d} = log_{10}\left(1 + tf_{t,d}\right) * (1 + log_{10}\,({}^{N}/_{df_t})$
4	Log-entropie = $\log\,(TF_{ij} + 1)$ *($\frac{\sum_j P_{ij.}\log(P_{ij})}{\log N}$)+1 with $P_{ij} = \frac{TF_{ij}}{\sum_j TF_{ij}}$
5	Log-entropie'= 0,8+0,8 * $\frac{TF_{ij}}{\max TF_{i'j}}$ * ($\frac{\sum_j P_{ij.}\log(P_{ij})}{\log N}$)+1 With $P_{ij} = \frac{TF_{ij}}{\sum_j TF_{ij}}$
6	Log-entropie''= 0,7+0,7 * $\frac{TF_{ij}}{\max TF_{i'j}}$ * ($\frac{\sum_j P_{ij.}\log(P_{ij})}{\log N}$)+1 with $P_{ij} = \frac{TF_{ij}}{\sum_j TF_{ij}}$
7	$TFIDF_{ij} = TF_{ij} * IDF_i$ with $IDF_i = \log \frac{N}{n_i}$
8	$TFIDF_{ij} = TF_{ij} * IDF_i$ Avec $TF_{ij} = \log\,(1 + TF_{ij}),\ IDF_i = \log \frac{N}{n_i}$
9	$BM25 = \sum idf * \frac{tf * (1+k)}{tf + k * (1 - b + b * \frac{D}{avgD})}$

We considered 2 categories of texts: a category that we named Text 1 containing long texts with an average of 2150 words, and a second category called Text 2 comprising short texts with an average of 500 words. The calculation of the f-measure was performed for each formula selected for the weighting of the occurrence matrix. We compared the results obtained by formula 1 used in our system to the other formulas. As shown in fig.3, Formula 1 shows an f-measure of up to 80% for Category 1 texts. The results are good thanks to the combination of syntax with the LSA and the extraction of terms with rigorous models.

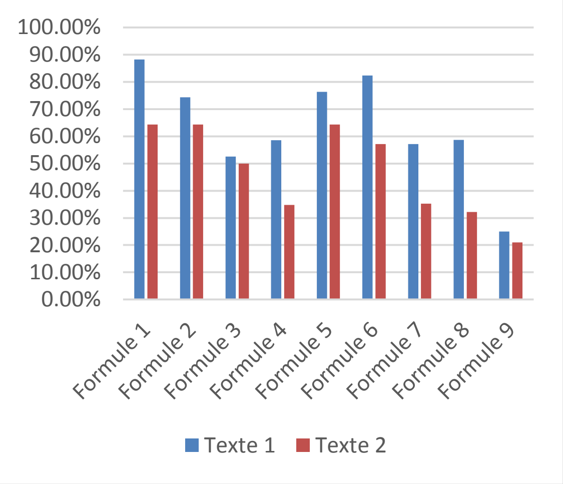

Figure 3. Graphical comparison of the f-measure for the different formulas on the two categories of texts

We also notice that f-measures of long texts are better than those of short texts because the LSA finds its performance only on large texts. Indeed, this approach is based on statistical measures, which are significant only on large documents.

To better illustrate this f-measure, take the example of a text composed of 578 words that we analyzed. Ar2Concept extracted 59 terms and generated 18 concepts. We note that in these 18 concepts, only 1 concept was not recognized and 3 of the concepts found were incorrect. Which is satisfactory. While for another text counting 370 words, of the 112 terms extracted, 13 concepts were identified among which 4 were incorrect and 5 unidentified.

These concepts not identified by Ar2Concept are mainly due to:

- A formula in itself that sometimes takes the frequency of the word in the text (importance is not necessarily due to the number of the appearance of the word in the text)
- An incorrect morphological analysis made by MADAMIRA. The authors speak of an error rate of 6%.
- Has a non-fine morphological analysis made by MADAMIRA, because this analyzer does not give a detailed label on the word, especially for nouns such as temporal and location adverbs which can be considered in some cases as concepts.
- At the very definition of the LSA because we can have two words that are not semantically close but they appear in similar contexts, which contradicts the first definition of LSA.

5. Conclusion

We achieved Ar2Concepts system. It's able to automatically extract concepts from texts written in Arabic and the results are more than satisfactory compared to all other extractor in Arabic. The method we adopted proved to be adaptable and we were able to do our experiments and evaluations on any corpus written in Arabic. The formula we named "BO19" that suggest to fix the value of k in the K-means algorithm has shown very good results. We wish to have contributed by this formula a solution to this problem. As perspectives, we will like to improve more the results obtained by our system and to add a module for the extraction of the relations between the previously extracted concepts in order to lead to ontology generation.

Acknowledgment

I would like to thank our ANLP-RG research team as well as our master students from the computer science department at the University Oran1 for their efforts in designing the Ar2Concept platform.

References

[1] Pouliquen, B. et Denis, D. Indexation de textes médicaux par extraction de concepts, et ses utilisations. Université Rennes 1, 2002.

[2] Bendaoud, R. Construction et enrichissement d'une ontologie à partir d'un corpus de textes. loria campus scientifique - bp 239 54506 vandoeuvre-lès-nancy cedex-lyon - france.2006.

[3] Mallak, I. De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information (Doctoral dissertation, Université Paul Sabatier-Toulouse III). 2011.

[4] Landauer, T. K., Foltz, P. W. et Laham, D. An introduction to latent semantic analysis. Discourse Processes, 25(2‑3), 259‑284. 1998.

[5] Abderrahim, M. A. Exploitation des Ontologies dans les Systèmes de recherche d'informations Arabes.. Repéré à http://dspace.univ-tlemcen.dz/handle/112/12634. Thesis, 13-03-2018.

[6] Bouaziz Mezghanni, I. et Gargouri, F. Towards an Arabic legal ontology based on documents properties extraction. Dans 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA) (p. 1‑8). Marrakech, Morocco : IEEE, 2015.

[7] Buitelaar, P., Cimiano, P., & Magnini, B. Ontology learning from text: An overview. Ontology learning from text: Methods, evaluation and applications, 123, 3-12.,2 005.

[8] Benaissa, B. E., Bouchiha, D., Zouaoui, A., & Doumi, N. Building ontology from texts. *Procedia Computer Science*, *73*, 7-15. 2015.

[9] A.Al-Arfaj and A. Al-Salman. Towards Concept Extraction for Ontologies on Arabic language. in Proceeding of International Journal on Islamic Applications in Computer Science And Technology, Vol. 4, Issue 4, pp 9 -19, December 2016.

[10] Dahab, M. Y., Idrees, A. M., Hassan, H. A., & Rafea, A. Pattern Based Concept Extraction for Arabic Documents. International journal of intelligent Computing and information sciences, 10(2), 1-14.2010.

[11] S. Zaidi., M. Laskri and A. Abdelali. Arabic collocations extraction using Gate. In Proceeding of International Conference on Machine and Web Intelligence (ICMWI). pp. 473 - 475. 2010

[12] A. Mashaan Abed., S. Tiun and M. AlBared. Arabic Term Extraction using Combined Approach on Islamic document. Journal of Theoretical & Applied Information Technology, 58 (3), pp.601-608,2013.

[13] A. El-Mahdaouy, S. Alaoui Ouatik and E. A study of association measures and their combination for Arabic MWT extraction. In Proceedings 11th International Conference on Terminology and Artificial Intelligence, pp. 45-52,2013.

[14] Benabdallah, A., Abderrahim, M. A., & Abderrahim, M. E. A. Extraction of terms and semantic relationships from Arabic texts for automatic construction of an ontology. International Journal of Speech Technology, 20(2), 289-296, 2017.

[15] Alarfaj, A., & Alsalamn, A. A New Concept Extraction Method for Ontology Construction From Arabic Text. International Journal of Artificial Intelligence and Expert Systems (IJAE), Volume (9) : Issue (1) : 2020

[16] J. Villalon and R. A. Calvo. Concept Extraction from Student Essays, Towards Concept Map Mining. Ninth IEEE International Conference on Advanced Learning Technologies, Riga, 2009, pp. 221-225, 2009

[17] Gong and X. Liu. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In International conference on Research and development in information retrieval, pp. 19-25, 2001.

[18] Gefen, D., Endicott, J. E., Fresneda, J. E., Miller, J., & Larsen, K. R. A Guide to Text Analysis with Latent Semantic Analysis in R with Annotated Code: Studying Online Reviews and the Stack Exchange Community. Communications of the Association for Information Systems, 41, 2017.

[19] Al-Sabahi, K., Zhang, Z., Long, J., & Alwesabi, K. An enhanced latent semantic analysis approach for arabic document summarization. *Arabian Journal for Science and Engineering*, *43*(12), 8079-8094. 2018.

[20] Imad Dabbura. K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Towards Data Science, a Medium publication sharing concepts, ideas, and codes.2018.

[21] Pasha, A., M., A.-B., Diab, M. T., El Kholy, A., Eskander, R., Habash, N. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In LREC ;Vol. 14, (pp. 1094-1101). 2014.

[22] Ghezaiel, L. B., Latiri, C., Ahmed, M. B. et Gouider-Khouja, N. Enrichissement d'ontologie par une base générique minimale de règles associatives. Laboratoire de recherche RIADI-GDL, ENSI, Campus Universitaire La Manouba, Tunis, 12. 2010.

[23] Pal, S.. Singular Value Decomposition for Recommendations with Tensorflow. Salmon Run. 5 mai, 2018. Repéré à https://sujitpal.blogspot.com/2018/05/singular-value-decomposition-for.html

Dr Bousmaha Ossoukine Kheira Zineb is a PhD of Computer Science at University Ahmed Benbella, Oran1 (Algeria). She teaches in Department of Computer Science and is member of Arabic Natural language Processing Research Group (ANLP-RG) and member of laboratory RIIR.

Pr Belguith Hadrich Lamia is a full Professor of Computer Science at Faculty of Economics and Management of Sfax (FSEGS) - University of Sfax (Tunisia). She teaches at the department of computer Science since 1992. She is Head of Arabic Natural language Processing Research Group (ANLP-RG) of Multimedia, InfoRmation systems and Advanced Computing Laboratory (MIRACL). She had published more than 220 papers.