Journal of Theoretical and Applied Information Technology
15th April 2012. Vol. 38 No.1
© 2005 - 2012 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ARABIC TEXT CATEGORIZATION: A COMPARATIVE STUDY OF DIFFERENT REPRESENTATION MODES

[1]ABIDI KARIMA, [2]EIBERRICHI ZAKARIA, [3]TLILI GUISSA YAMINA

[1] Ecole supérieur d'informatique, Algeria

[2] EEDIS Laboratory, UDL, Sidi Belabbes. Algeria

[3] Université Badji Mokhtar-Annaba, Alegria

E-mail: [1]k_abidi@esi.dz , [2] elberrichi@univ-sba.dz, [3] guiyam@yahoo.fr

ABSTRACT

The quantity of accessible information on Internet is phenomenal, and its categorization remains one of the most important problems. A lot of work is currently focused on English rightly since; it is the dominant language of the Web. However, a need arises for the other languages, because the Web is each day more multilingual. The need is much more pressing for the Arabic language. Our research is on the categorization of the Arabic texts, its originality relates to the use of a conceptual representation of the text. For that we will use Arabic WordNet (AWN) as a lexical and semantic resource. To comprehend its effect, we incorporate it in a comparative study with the other usual modes of representation (bag of words and N-grams), and we use different similarity measures. The results show the benefits and advantages of this representation compared to the more conventional methods, and demonstrate that the addition of the semantic dimension is one of the most promising approaches for the automatic categorization of Arabic texts.

Keywords: *Categorization, KPPV, Arabic wordnet ,* N-grams .

1. INTRODUCTION

The emergence of the Internet, the enormous increase in the amount of information and digital resources on one side and the globalization of the world on the other hand, has changed deeply the means of communication, in particular, by facilitating the exchanges of documents between different cultures and countries, and thus created new needs for users to exploit this wealth of information. Among these needs, the improvement of how to find relevant information in a language other than English. Lately a growing interest is specifically on the collections of information written in Arabic.

Text categorization is to find a functional link between a set of texts and a set of categories (labels, classes). This functional relation, also called the prediction model is estimated by a supervised learning system. To do this, it is necessary to have a set of previously labeled texts, called the learning set, from which are estimated the model parameters for the best possible prediction [4].

Research on categorization in other languages, and English in particular, has shown that the performance of a system depended largely on how the text is represented and the similarity measure used besides the learning algorithm. This is why it would be coherent in a comparative study that the most used representation and the measurements most present are tested.

To identify the category or class that is associated with a text, a set of steps is usually followed. These steps are primarily a preprocessing on the text, a choice on the mode of representation and a reduction of the dimensionality. The categorization process involves two phases: the learning phase and the classification phase.

2. RELATED WORK

Compared with other languages, there is little research on the classification of documents in Arabic. Harbi compared the two learning algorithms C5.0 and SVM with the bag of words representation and concluded on the superiority of C5.0 [2]. Khreisat's results show that N-gram text classification using the Dice measure outperforms classification using the Manhattan measure for a corpus unfortunately limited to 4 categories [7]. Sawaf used the maximum entropy for Arabic text classification and has shown that statistical methods are promising, even without any morphological analysis [4]. In [5] El-Kourdi has used the Naive Bayes algorithm, the reported results were

© 2005 - 2012 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

interesting. Al-Shalabi [6] improved the K-NN algorithm by using the n-gram representation.

3. ARABIC TEXT CATEGORIZATION

In this work, we used a corpus of Arabic texts built by Mesleh [8]. from online newspapers such as Al-Jazeera, Al - Nahar, Al-Hayat and Al-Dostor, and some other specialized sites. This corpus is widely used for the experiments of text mining applications on the Arabic language. It consists of 1445 documents (14 MB). This corpus is written in the standard Arab language except for some religious texts (Al Hadith and Qur'an) which contain a mixture of the classical and modern arabic.
It consists of 1445 documents (14 MB) classified into nine categories. We split the corpus as follows: 66% of all texts for learning and the rest for testing approaches of representation. This gave 965 texts for learning and 480 texts for testing or evaluation.

3.1. The text preprocessing

This is an important phase in the learning process. It is necessary to clean the texts by removing stop words "(articles, determinants, auxiliaries... etc), the words which are of lower value to the text. We implement this phase in three steps:

3.1.1. Single text encoding: The encoding of texts in one standard format is used to represent texts without any deformation of character during the reading. . All our corpus texts are represented with an ANSI encoding, the encoding supported by the java language.

3.1.2. Removing stop words: It consists in eliminating all no significant words belonging to the stop words list constructed on the basis of the list of **Kadr** and **Khoja** [6] and completed by our care where we have added other missing preposition as « أيضا, التي ...etc. », and in Arabic the pronoun is often glued to a preposition we can generate a multitude of forms for stop word if pasting the following pronouns with antefixes« بـ,ل,ف,و,ك» or postfixe(هن,هما,هم..etc). We also have removed the non-Arabic letters, digits, single Arabic letters (letters that dose not belong to word), punctuations, special symbols, diacritics«ً, ٌ, ٍ, َ, ُ, ِ, ّ, ْ ». word punctuation marks and also kashida character (add the line in the middle of the Arabic word(منزل مـنـــــزل)) are also eliminated.

3.1.3 Morphological processing: this step specific for the Arabic language. We applied a normalization morphological of some characters [6][10], This method based on linguistic concepts tries to determine the core of a word according to linguistic rules conformed by statistics as follows:

- Replacing أ and إ by ا.
- Removing ALF-tanwin أ
- Replacing ة by ه
- Removing وال, فال , بال , and كال
- Removing ال , and لل except from الله .

3.2 The different modes of representations

There are three important text representation modes in text mining. In our experiments, we decided to experiment on these three representations to compare them:

3.2.1. The representation "Bag of words":

The simplest representation of texts introduced within the framework of the vectorial model. The idea is to transform the texts into vectors of words. This representation excludes any form of grammatical analysis and any notion of distance between words [12].

3.2.2 The representation based on the N-grams:

An N-gram is a sequence of N characters. In this paper, an N-gram will indicate a chain of N consecutive characters. In the literature, this term refers sometimes to sequences that are not ordered or consecutive. For any document, all N-grams (in general N takes values 2.3 or 4) generated are the result obtained by moving a window of N boxes on the text body. This displacement is done by steps; a step corresponds to a character. Then the frequencies of the found N-grams are counted. For the N-representation, the removing of stop words and the morphological normalization are usually not necessary but can improve the results [9].

3.2.1 Representation based on concepts:

While also relying on the vector formalism to represent documents, the vector elements are no longer directly associated to the text terms but to the text concepts.
To allow such a representation of documents, it is necessary to be able to project our terms on a thesaurus or a lexicon such as WorldNet [3]. WordNet is a lexical reference system whose design is strongly inspired by the recent psycholinguistiques theories on human semantic memory. The nouns, verbs, adjectives and adverbs in English are organized into sets of synonyms

Journal of Theoretical and Applied Information Technology
15th April 2012. Vol. 38 No.1
© 2005 - 2012 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(synsets), each one representing a lexical concept. Various relations bind the synsets in a semantic network. For this representation, two approaches are possible, either the vector of representation contains only the concepts associated with the words of the text represented, or it contains and the concepts and the words [11].

Arabic WordNet contains 9228 concepts or synsets (6252 nominal, 2260 verbal, 606 adjectival, and 106 adverbales), 18.957 expressions and 1155 named concept [15].

The principle of this approach is to find the concept that represents a set of synonyms. opt for this approach we must choose the appropriate concept from the concepts found. in our work for each word we selected the concept that contains the maximum of the words in the representative vectors.

For example the word كتبin (**Figure** 1) belongs to 12 concepts but only up to the 12th contains synonyms listed in the representative vector the word (دون) represents the concept and the new frequency IDF is the sum of the frequencies of all frequencies synonyms found in the representative vector.

N°	Le concept Arabe
01	اشترك اكتتب استلم شرك كتب سلم
02	دون كتب سجل
03	كتب كاتب
04	كاتب مؤلف كتب ألف
05	ألف كتب نظم
06	كتاب مجلد كتب جلد
07	كتابه تاليف تحرير كتب ألف حرر
08	كتابه تحرير كتب حرر
09	كتابي تحريري كتب حرر
10	مكتب موظفين كتب وظف
11	كتابي مكتوب مخطوط مؤلف مدون كتب كتب خطط ألف دون
12	دون كتب قيد سجل وضعَ

كتب	قيد	رسم	سجل	دون
2	4	1	7	3

Representative vector of each category

Wordnet Concept **كتب** { دون كتب قيد سجل وضع دَوَّنَ كَتَبَ قَيَّدَ سَجَّلَ وَضَعَ}
New frequency 16

Figure 1: The procedure for grouping words.

3.3The reduction of the dimensionality

Our corpus is very large, as it is usually the case of text categorization application (the curse of dimensionality). We must therefore reduce the size of the vectors; we use the Khi2 method for selecting only the most representative terms.

	C	Not C	total
T	A	B	A+B
Not T	C	D	C+D
Total	A+C	B+D	N

$$X^2(t,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

Where

N = total number of documents in the corpus.
A = Number of documents in class C that contain the term t.
B = Number of documents containing the term t in other classes.
C = Number of documents in category C, which does not contain the term t.
D = Number of documents that do not contain the term t in other classes.

3.4Algorithm K nearest neighbors (K-NN)

K-NN is an algorithm that has proven its effectiveness in the supervised classification of textual data. The learning phase consists in storing the labeled examples (vectors representing the texts and their class). The classification of new texts is made by calculating the distance between the vector representing the document and each stored instance of the corpus. The K Nearest instances are selected and the document is assigned the majority class (the weight of each class may be weighted according to its distance). A variant of this method is used for automatic classification [11]. For a comparative study as complete as possible, and because the similarity measure plays a crucial role in the method, we used the three similarity measures mentioned below:

a. The jaccard measure:

$$SIMJaccard(d_{ik}, q_k) = \frac{\sum_{k=1}^{t}(d_{ik} \bullet q_k)}{\sum_{k=1}^{t} d_{ik}^{\ 2} + \sum_{k=1}^{t} q_k^{\ 2} - \sum_{k=1}^{t}(d_{ik} \bullet q_k)} \quad (1)$$

b. The cosinus measure:

$$SIM\ \text{cosinus}\quad (d_{ik}, q_k) = \frac{\sum_{k=1}^{t}(d_{ik} \bullet q_k)}{\sqrt{\sum_{k=1}^{t} d_{ik}^{\ 2} \bullet \sum_{k=1}^{t} q_k^{\ 2}}} \quad (2)$$

c. The inner measure:

$$SIMInner(d_{ik}, q_k) = \sum_{k=1}^{t}(d_{ik} \bullet q_k) \quad (3)$$

Journal of Theoretical and Applied Information Technology
15th April 2012. Vol. 38 No.1
© 2005 - 2012 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

d. The Dice measure:

$$SIMDice\ (d_{ik}, q_k) = \frac{2 * \sum_{k=1}^{t} (d_{ik} \bullet q_k)}{\sum_{k=1}^{t} d_{ik}^{\ 2} + \sum_{k=1}^{t} q_k^{\ 2}} \quad (4)$$

4. RESULTS:

We have built text classification system which in reality includes three classifiers: the first is based on single terms "Bag of words" the second using N-Grams as documents while the third used concept. We tested our sytem using the test set which is about 34% from the data set size and for each categories precision and recall is calculated for each method. based on presision and recall, the F-mesure is calculated(Figure2,Figure3, Figure4).
The average **Precision**, **recall** and **F**-measure is calculated as follows:

$$\Pr ecision = \frac{\#correct - classes - found}{\#classes - found} \quad (5)$$

$$\mathrm{Re}call = \frac{\#correct - classes - found}{\#correct - classes} \quad (6)$$

$$F - Measure = \frac{2.P.R}{P + R} \quad (7)$$

The results in the **Figures** (Figure2,Figure3, Figure4). Represent the F-measure for each approach allows us to note that the cosine measure is the most effective for a classifier based on the algorithm KPPV, and also show that the use of external resources in the process of categorization as Arabic WordNet is a very promising.

Results of the approach Bag of words:

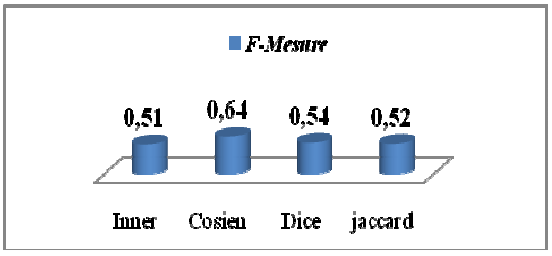

Figure 2: The recall, precision and F-measure for the bag of words approach.

Results of the N-gram approach:

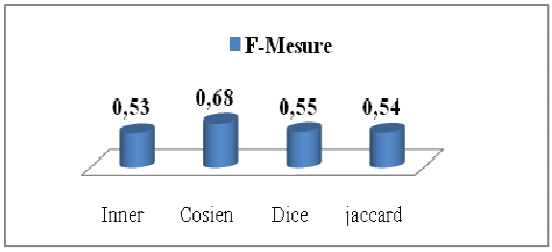

Figure 3: The recall, precision and F-measure for the N-gram approach (tri-gram).

Results of the conceptual approach:

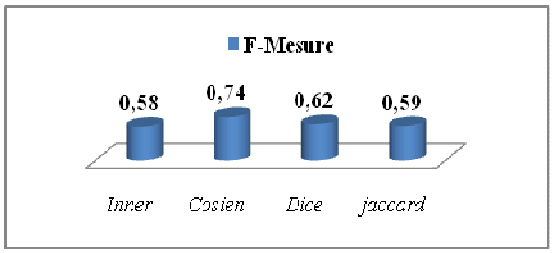

Figure 4: The recall, precision and F-measure for the approach based on the concept.

5. CONCLUSION

To the best of our knowledge, we are the first to propose a conceptual representation for arabic text representation. For that we used AWN to map the terms to concept, this is also a first. Its counterpart WordNet has been widely used for that purpose and others, specially for text mining applications. We suppose the same thing is going to happen to AWN. But most important, we think that bringing the semantic dimension to the classification of Arabic text and any other Arabic text mining application is a very promising approach. Our results for classification have demonstrated that. For the future we hope to demonstrate that on other applications.

in our future work we will use Arabic WordNet to enrich the learning of word semantic orientation to find documents

© 2005 - 2012 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFRENCES:

[1] T.S. Bhatti, R.C. Bansal, and D.P. Kothari, "Reactive Power Control of Isolated Hybrid Power Systems", *Proceedings of International Conference on Computer Application in Electrical Engineering Recent Advances (CERA)*, Indian Institute of Technology Roorkee (India), February 21-23, 2002, pp. 626-632.

[1] Al-Harbi, S., Almuhareb,A., Al-Thubaity, A. Khorsheed, M.S., Al-Rajeh,A., *« Automatic Arabic Text Classification. »* JADT 2008: 9ème Journée Internationales d'Analyse Statistique des Données Textuelles.

[3] Simon JAILLET 2004, *Catégorisation automatique de documents.*

[4] Jalam Radwan, (2003) "Apprentissage automatique et catégorisation de textes multilingues". Thèse de doctorat, Université Lumière Lyon 2.

[5] Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran, Musa Alkhalifa, M. Antonia Martí(2008) « Arabic WordNet: Semi-automatic Extensions using Bayesian » Inference Proceedings of the the 6th Conference on Language Resources and Evaluation LREC2008.

[6] Kadri Youssef, Thèse présentée à la Faculté des études supérieures en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.) en informatique septembre 2008 «Recherche d'Information Translinguistique sur les Documents en Arabe »

[7] Khreisat Laila, « Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study », In : Proceedings of the 2006 International Conference on Data Mining

[8] MESLEH Abdelwadood Moh'd , 2007 « Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System » Faculty of Information Systems and Technology, Arab Academy for Banking and Financial Sciences, Amman, Jordan.

[9] Nicolas BÉCHET, thèse de doctorat Soutenue le 8 décembre 2008 « Extraction et regroupement de descripteurs morphosyntaxiques pour des processus de Fouille de Textes » présentée à l'Université des Sciences et Techniques du Languedoc pour obtenir le diplôme de DOCTORAT

[10] Rasha Obeidat , Riyad Al-Shalabi, Improving KNN Arabic Text Classification with N-Grams Based Document Indexing

[11] William J. Christiane Fellbaum, Musa Alkhalifa, Black, Sabri Elkateb, Adam Pease, Horacio Rodríguez, Piek Vossen (2006) « Introducing the Arabic WordNet project Proceedings of the 3rd Global Wordnet Conference », Jeju Island, Korea, January, 2006.

[12] Young-Min Kim, Jean-François Pessiot, Massih-Reza Amini, Patrick Gallinari (2008) « Apprentissage d'un espace de concepts de mots pour une nouvellereprésentation des données textuelles

[13] Sawaf, H., Zaplo,J., Ney, H. « Statistical classification methods for Arabic news articles. Arabic Natural Language » Processing Workshop (ACL 2001); Toulouse, France .

[14] El-Kouridi M, Bensaid A.et Rachidi T,2004 automatic Arabic Association Rules for texte classification .InProcesdings of the firste intrentional

[15] Christiane Fellbaum, Musa Alkhalifa, William J. Black, Sabri Elkateb, Adam Pease, Horacio Rodríguez, Piek Vossen (2006) « Introducing the Arabic WordNet project Proceedings of the 3rd Global Wordnet Conference », Jeju Island, Korea, January, 2006