Arabic Text Categorization: a Comparative Study of Different Representation Modes

Zakaria Elberrichi and Karima Abidi
Department of Computer Science, Djillali Liabes University, Algeria

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 the K-Nearest Neighbors (K-NN) learning scheme with 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 ways for the automatic categorization of Arabic texts.*

Keywords: *Categorisation, Arabic texts, AWN, bag of words, ngrams, concepts.*

Received May 27, 2010; accepted August 10, 2010

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.

Formally, the text categorization is to assign a Boolean value to each pair $(dj, ci) \in D \times C$, where D is the set of texts and C is the set of categories. The value True (T) is then associated to a pair (d_j, c_i) if the text d_j belongs to the class c_i while the latter value False (F) will associate it otherwise. The goal of text categorization is to construct a procedure (model, classifier) $\Phi: D \times C \rightarrow (T, F)$ which associates one or more categories with a document d_j such as the decision given by this procedure "coincides as much as possible" with function $\Phi: D \times C \rightarrow \{T, F\}$, the true function which turns over for each vector d_j a value C_i [14]. Thus, 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. Research on categorization in other languages, and English in particular, has shown that the performance of a model depended largely on how the text is represented and on the similarity measure used besides the learning algorithm [4]. This is why it would be coherent in a comparative study that the most used representation and the measurements most present are tested. Before identifying 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 of the mode of representation and a reduction of the dimensionality. The categorization process involves two phases: the learning phase and the classification phase. Figures 1 and 2 show graphically the general outline of a supervised text classification. The novelty is that the text is in Arabic. This characteristic will require some special considerations that will be addressed later in the following section. Section 2 presents some related works, section 3 introduces the corpus we used to test our approaches and the preprocessing done to the text, section 4 details the three representation modes, section 5 deals with the dimensionality issue, section six is about the learning algorithm K-Nearest Neighbors (K-NN) used, the last sections presents as usual the results and the conclusion.

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 [1]. 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 [11]. Sawaf used the maximum entropy for Arabic text classification and has shown that statistical methods are promising, even without any morphological analysis [13]. El-Kourdi *et al*. [6] has used the Naive Bayes algorithm, the reported results were interesting. Al-Shalabi and obeidat [3] improved the K-NN algorithm by using the n-gram representation. Al-Kabi and Al-Sinjilawi [2] did a comparative study of the efficiency of different measures to classify Arabic text. The experiments related to this study were conducted on a set of Arabic text documents, where each document is a Hadith (saying) of the Prophet Mohammed (PBUH). The results show that Naïve Bayesian is the first best, the second is Cosine.

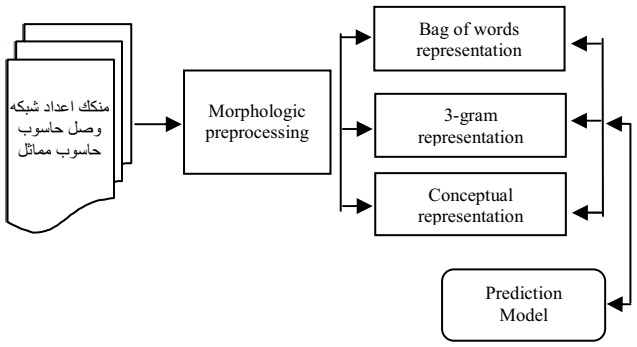

Figure 1. The learning phase.

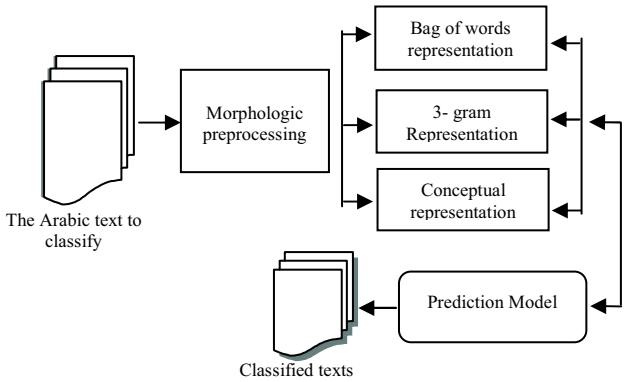

Figure 2. The classification phase.

3. The Corpus and the Preprocessing

3.1. The Corpus

In this work, we used a corpus of Arabic texts built by Mesleh [12] from online newspapers such as Al-azeera, 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 (14MB) of variable length (between 1KB and 106KB or 30 words and 19072 words) and classified in nine categories. This corpus is written in the standard Arabic language except for some religious texts (Al Hadith and Qur'an) which contain a mixture of the classical and modern Arabic. Table 1 details their distribution. We split the corpus as follows: 66% of all texts for training and the rest to test our approaches. It is generally, a tradition very followed by the researchers of the field. This gave 965 texts for training and 480 texts for testing or evaluation. A sample text of the computer category:

ولا يتطلب هذا الإعداد Peer-To-Peer Network حاسوب آخر مماثل لحاسوب مزودمكرس للشبكة إو إدارة علي مدار الساعة . ويمكن للحواسيب الموصولة بشبكة من هذا النوع المشاركة بالموارد والطابعات وأجهزة المودم وسواقات الأقراص المدمجة وسعات الأقراص الصلبة، وتتيح كل إصدارات ويندوز المشاركة بهذه الموارد، ومثلا يمكن لكل الحواسيب الموصولة بشبكة حواسيب متماثلة استخدام طابعة موصولة بحاسوبك إذا منحت امتياز الوصول

Figure 3. A sample of an Arabic text.

Table 1. MESLEH's Arabic texts corpus.

Categories	Number of Documents
Computer	70
Economic	220
Engineer	115
Education	68
Law	97
Politics	184
Religion	227
Sport	232
Medicine	232
Total	1445

3.2. 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:

- *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.
- *Removing Stop Words:* It consists in eliminating all nonsignificant words by removing stop words (articles, determinants, auxiliaries... etc.,), the words which are of lower value to the text and belonging to the stop word's list conceived on the basis of the list of [2, 10] and completed by our care. Punctuation marks, numbers, words in Latin characters, abbreviations and single letters are also eliminated.
- *Morphological Processing:* Specific for the Arabic language. In addition to morphological standardization of some characters, these processing primarily affect the lemmatization of Arabic words. A method of lemmatization is implemented. This method based on linguistic concepts tries to

determine the core of a word according to linguistic rules conformed by statistics as follows:

- Delete Alef El Tanwin "ً".
- Replace all "إ", "أ", "آ" by "ا".
- Replace all "ى" by "ي".
- Replace all "ة" by "ه".

4. 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.

4.1. 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.

Table 2. Number of words in each category.

Category	Number of Words
Computer	9679
Economic	93888
Engineer	94340
Education	43753
Law	95805
Politics	59410
Religion	123111
Sport	49101
Medicine	65012
Total	634099

4.2. 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.

Table 3. Number of trigrams in each category.

Category	Nbr of 3-Ngram	Nbr of 3-Gram without Repetition
Computer	54086	4767
Economic	549093	8942
Engineer	518001	9084
Education	243457	8468
Law	544287	8482
Politics	343970	8809
Religion	657537	10092
Sport	279537	6479
Medicine	358959	8555
Total	3548927	73678

4.3. 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 [9]. WordNet [8] is a lexical reference system whose design is strongly inspired by the recent psycholinguistics theories on human semantic memory. The nouns, verbs, adjectives and adverbs in English are organized into sets of synonyms (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.

Arabic WordNet (AWN) [8] is a lexical resource for standard modern Arabic based on Princeton WordNet and is built according to methods developed for EuroWordNet. AWN can be related to other WN of other languages, allowing for translation from and into tens of languages. The connection of WordNet to Suggested Upper Merged Ontology (SUMO) is also an asset.

AWN contains 9228 concepts or synsets (6252 nominal, 2260 verbal, 606 adjectival, and 106 adverbials), 18957 expressions and 1155 named concept. The file bases AWN under XML format contains the four tags:

- *Item:* Contains (synsets) concepts, classes, and instances of the ontology.
- *Word:* Contains words.
- *Form:* contains the roots of Arabic words.
- *Link:* Relationships between concepts.

Our work consists in for this stage representing the similar words (synonyms) in one concept.

Table 4. Number of concepts compared to Number of words in each category.

Category	Number of Words	Number of Concepts
Computer	2521	281
Economic	13798	745
Engineer	13404	765
Education	11018	606
Law	12959	200
Politics	12000	299
Religion	17304	757
Sport	5079	427
Medicine	10363	681

Note: In AWN, the words come with el tchekile and even the roots of the words, and since in most of the modern Arabic texts, texts are written without techkile and since in the cleaning phase, we removed el tchekile, we opted to do the same thing with the word concepts selected; we converted them into a standard

format. This is a non perfect intermediate solution, but generally an adopted one.

Figure 4. Example concepts possible for the word وصل in the browser of AWN [10].

In our approach we recommend the replacement of terms by concepts, and we use a simple disambiguation strategy when, there is more than one concept proposed. We take the first concept as the most suitable. WorldNet gives for each term a list of ordered concepts arranged from the most appropriate concept to the less appropriate. The frequency of a concept is then calculated as follows:

$$cf(d,c) = tf\{d, \in T \mid first(ref,(t)) = C\}\} \quad (1)$$

5. The Reduction of the Dimensionality

Our corpus is very large, as it is usually the case for text categorization application (the curse of dimensionality). The use of the concepts instead of the words reduces considerably the dimensionality as Table 4 has shown it. To reduce further the size of the vectors; we use the Khi2 method for selecting only the most representative terms.

Table 5. Confusion matrix.

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

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

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.

6. Algorithm 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 new 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). 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:

1. *Cosine Similarity:* Is a measure of similarity between two vectors of *n* dimensions by finding the cosine of the angle between them. Given two vectors of attributes, *A* and *B*, the cosine similarity is represented using a dot product and magnitude:

$$similarity = \cos(\theta) = \frac{A.B}{\| A \| \| B \|}$$

2. *The Jaccard Similarity Coefficient:* Is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$similarity = \frac{|A \cap B|}{|A \cup B|}$$

3. *Finally, The Inner Product Similarity:* Also known as the dot product or scalar product:

$$similarity = A.B$$

7. Results

Text categorization effectiveness is measured in terms of precision and recall [5]. Precision and Recall are defined as [15]:

$$Recall = \frac{a}{a+c} \; a+c > 0 \qquad Precision = \frac{a}{a+c} \; a+b > 0$$

where *a* counts the assigned and correct cases, *b* counts the assigned and incorrect cases, *c* counts the not assigned but incorrect cases and *d* counts the not assigned and correct cases.

The values of precision and recall often depend on parameter tuning; there is a trade-off between them.
This is why we also use another measure that combine both of the precision and recall: the F-measure which is defined as follows:

$$F\text{-}measure = \frac{2(Precision \times Recall)}{Precision + Recall}$$

To evaluate the performance across categories, F-measure is averaged. The results of the approaches are shown in the following Tables 6, 7, and 8, and the best macro-averaged F-measures are underlined. It is clear that the use of concepts is a better way of representing Arabic texts, the performances as shown by the F-measure are better and that's true with the three distance measures. Of the three distances, the cosine distance performs the best. That's what we expected [7].

Table 6. Results of the approach Bag of words.

Bag of Words	Inner	Cosine	Jaccard
Rappel	0.47	0.63	0.47
Precision	0.69	0.73	0.70
F-measure	0.55	**0.67**	0.56

Table 7. Results of the N-gram approach.

N-gram	Inner	Cosine	Jaccard
Recall	0.46	0.61	0.46
Precision	0.70	0.77	0.73
F-measure	0.55	**0.68**	0.57

Table 8. Results of the conceptual approach.

Conceptual	Inner	Cosine	Jaccard
Recall	0.53	0.67	0.54
Precision	0.72	0.77	0.73
Le F-measure	0.60	**0.74**	0.61

8. Conclusions

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 [7] and others, especially 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 of arabic have demonstrated that. For the future we hope to demonstrate that on other applications.

References

[1] Al-Harbi S., Almuhareb A., Al-Thubaity A., Khorsheed M., and Al-Rajeh A., "Automatic Arabic Text Classification," *in Proceedings of The 9th International Conference on The Statistical Analysis of Textual Data*, France, pp. 77-83, 2008.

[2] Al-Kabi M. and AlSinjalawi S., "A comparative Study of the Efficiency of Different Measures to Classify Arabic Text," *University of Sharjah Journal of Pure & Applied Sciences*, vol. 4, no. 2, pp. 13-24, 2007.

[3] Al-Shalabi R. and Obeidat R., "Improving KNN Arabic Text Classification with N-Grams Based Document Indexing," *in Proceedings of the 6th International Conference on Informatics and Systems*, Egypt, pp. 108-112, 2008.

[4] Amine A., Elberrichi Z., and Simonet M., "Evaluation of Text Clustering Methods Using WordNet," *The International Arab Journal of Information Technology*, vol. 7, no. 4, pp. 349-357, 2010.

[5] Baeza-Yates R. and Rieiro-Neto B., *Modern Information Retrieval*, ACM Prss/Addison-Wesley, 1999.

[6] El-Kourdi M., Bensaid A., and Achidi T., "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," *in Proceedings of the Workshop on Computational Approaches to Arabic Script Based Languages*, Geneva, pp. 51-58, 2004.

[7] Elberrichi Z. and Rahmoun A., "Using Wordnet for Text Categorization," *The International Arab Journal of Information Technology*, vol. 5, no. 1, pp. 16-24, 2008.

[8] The Global Wordnet Association, available at: http://www.globalwordnet.org/, last visited 2012.

[9] Jaillet S., Catégorisation Automatique de Documents LIRMM UMR, available at: http://www.lirmm.fr/doctiss04/art/I02.pdf, last visited 2004.

[10] Khodja S. and Garside S., "Stemming Arabic Text," *Technical report*, Computing Department, Lancaster University, UK, 1999.

[11] Khreisat L., "Arabic Text Classification Using N-Gram Frequency Statistics a Comparative Study," *in Proceedings of the International Conference on Data Mining*, USA, pp. 78-82, 2006.

[12] Mesleh M., "Chi Square Feature Extraction Based SVMs Arabic Language Text Categorization System," *Journal of Computer Science,* vol. 3, no. 6, pp.430-435, 2007.

[13] Sawaf H., Zaplo J., and Ney H., "Statistical Classification methods for Arabic News Articles," *in Proceedings of Arabic Natural Language Processing workshop*, France, pp. 1-6, 2001.

[14] Sebastiani F., "A Tutorial on Automated Text Categorisation," *in Proceedings of 1st Argentinian Symposium on Artificial Intelligence,* Buenos Aires, pp. 7-35, 1999.

[15] Yang Y., "An Evaluation of Statistical a Pproaches to Text Categorization," *Journal of Information Retrieval*, vol. 1, no. 1-2, pp. 69-90, 1999.

Zakaria Elberrichi received his Master degree in computer science from the California State University, in addition to PGCert in higher education, and received his PhD in computer science from the university Djillali Liabes, Sidi-Belabbes, where he has been a faculty member ever since. He is also a member of Evolutionary Engineering and Distributed Information Systems (EEDIS) laboratory and the project head-manager of the intelligent web mining team.

Karima Abidi received her Master degree from Ecole Superieure d'Informatique, Algiers. She is currently a Doctorate student and member of the research team intelligent web mining, working currently on a project called semantic web mining.