An Arabic Lemma-Based Stemmer for Latent Topic Modeling

Abderrezak Brahmi[1], Ahmed Ech-Cherif[2], and Abdelkader Benyettou[2]
[1]Department of Computer Sciences, Abdelhamid Ibn Badis University, Algeria
[2]Department of Computer Sciences, USTO-MB University, Algeria

Abstract: *Developments in Arabic information retrieval did not follow the increasing use of the Arabic Web during the last decade. Semantic indexing in a language with high inflectional morphology, such as Arabic, is not a trivial task and requires a text analysis in the original language. Excepting cross-language retrieval methods or limited studies, the main efforts, for developing semantic analysis methods and topic modeling, did not include Arabic text. This paper describes our approach for analyzing semantics in Arabic texts. A new lemma-based stemmer is developed and compared to root-based one for characterizing Arabic text. The Latent Dirichlet Allocation (LDA) model is adapted to extract Arabic latent topics from various real-world corpora. In addition to the interesting subjects discovered in the press articles during the 2007-2009 period, experiments show that the classification performances with lemma-based stemming in the topics space, are improved when comparing to classification with root-based stemming.*

Keywords: *Arabic stemming, topic model, semantic analysis, classification, test collection.*

Received October 22, 2010; accepted May 24, 2011

1. Introduction

Arabic is one of the top ten languages in the Internet. For a global population of 350 millions in Arabic world, Internet World Stats[1] have reported, for Internet Arabic users, the highest growth rate with 2,501.2% for the period 2000-2010. However, developments in Arabic information retrieval, and in other non-English languages, did not follow this extraordinary growth. In fact, English is used as the basis for most corpora in the different IR tasks such as search, document clustering and information extraction. In the tasks relating to semantic analysis, it is preferable to directly deal with text in its original language. Apply the same algorithms on parallel corpora by using English as the pivot language can alter the evaluation of these methods for the languages with high inflectional morphology such as Arabic [15, 19].

In this context, three levels of difficulties face the developments of the Arabic IR and generalization of new methods on Arabic texts: 1). How to efficiently extract a good stem when a morpheme implies several segmentations and senses? 2). How one can apply topic models to handle the semantic embedded in Arabic texts? 3). How to make more accessible Arabic resources for IR tasks to benefit from the various developments in non-commercial context.

This work is aimed at answering the two first questions raised above. On the one hand, two Arabic stemming methods will be tested as pre-treatment on Arabic newspaper articles. On the other hand, the Latent Dirichlet Allocation (LDA) model is adapted to extract Arabic topics from real world corpora. Concerning the third question, we program to regularly publish the works carried out on the University official website.

This paper presents firstly related works about Arabic IR tasks and topic modeling. Then, the nature of Arabic language and approaches for its text analysis are described. The generative process of LDA is illustrated before presenting the three datasets of newspaper articles automatically crawled from the Web. After detailing the experimentations completed and the results arrived at, the conclusion shows the underlying main observations of this work.

2. Related Works

2.1. Challenges in Arabic IR

For Arabic information retrieval, several light stemmers based on heuristics have been developed [14]. Among eight stemmer variants applied in the TREC-10 AFP_ARB corpus, the authors have found that for cross-language retrieval, light stemming was more effective than morphological analysis. They deduce that it is not essential for a stemmer to yield the correct form. Surprisingly, the authors claim, in a technical report, that these results were obtained with no prior experience with Arabic.

[1]Internet World Stats, usage and population statistics (Miniwatts Marketing Group). Updated for June 30, 2010, last visited in August 2010. http://www.internetworldstats.com/.

It is worth pointing out that the AFP_ARB corpus is one of the few standard Arabic datasets which consists of hundreds of thousands of Arabic newspaper articles collected from "Agence France Presse". The Linguistic Data Consortium has published it with license fees for non-members.

In document classification, few works for Arabic texts have been identified. For classifying 1,500 Web documents equitably divided in five categories, Naïve Bayes algorithm has been used in [8]. An accuracy of 68.78% has been reported. Maximum entropy method has been applied to classify, in six categories, real-world data collected from several Arabic Websites [7]. When using AraMorph[2] package for stemming task with additional preprocessing techniques, the author has increased the f-measure from 68.13% to 80.41%.

It is worth pointing out that the main efforts to build efficient IR systems for the Arabic language have been established in a commercial framework and the approach used as well as the accuracy of the performance are not known. As a significant example, Siraj system[3] from Sakhr allows to classify Arabic text and to extract named entities with human-satisfying response. However, it has no technical documentation to explain the used method neither the system evaluation.

2.2. Topic Modeling

For Arabic topic modeling, few studies were identified. In addition to some works related to Arabic topic detecting and tracking [15, 19], a segmentation method based on PLSA model [9] were applied to AFP_ARB corpus for monolingual Arabic topic analysis [3].

In [15], the authors claimed that it should be preferable to build separate language specific topic models for comparing different topic tracking methods. Good topic models have been obtained when native Arabic stories are available. However, Arabic topic tracking has not been improved when translating from English stories.

The LDA model has been introduced within a general Bayesian framework where the authors developed a variational method and EM algorithm for learning the model from collection of discrete data [2]. The authors applied their model in English document modeling, text classification and collaborative filtering.

Since the original introduction of LDA, several contributions have been proposed according to four directions: 1). Improving learning algorithm [17, 20], 2). Dealing with high scalability [18], 3). Developing topic model variants [6, 17], 4). Visualizing and interpreting results [10].

It is only recently that studies have begun to provide a comprehensive framework for LDA to deal with multilingual latent topics. Indeed, a method, for extracting multilingual topics from unaligned corpora, has been proposed and applied in English-Spanish and English-German parallel datasets [11]. For extracting a set of common latent topics from English-Chinese datasets, a new study has proposed the Probabilistic Cross-Lingual Latent Semantic Analysis as an extension of PLSA model [24].

3. Arabic Text Analysis

Unlike the indo-European languages, Arabic belongs to the Semitic family of languages. Written from right to left, it includes 28 letters. Despite the fact that different Arabic dialects are spoken in the Arab world, there is only one form of the written language found in printed works which it is known as the Modern Standard Arabic, herein referred to as Arabic [12]. The combination of the agglutination and the non-vocalization in Arabic text make a fundamental divergence with other languages, are.

3.1. Arabic Language Features

Arabic is a highly inflected language due to its complex morphology. An Arabic word can be one of three morpho-syntactic categories: nouns, verbs or particles. Several works have used other categories (such as preposition and adverb) with no good reason except that they are taken from English [14, 16, 22].

In Semitic languages, the root is an essential element from which various words may be derived according to specific patterns or schemes. Table 1 gives some simple derivations from the root [Elm] علم. The root is a linguistic unit carrying a semantic area. It is a non-vocalized word and often consists of only 3 consonants (rarely 4 or 5 consonants) [12, 22].

Table 1. Some derivations from the [Elm] علم root.

Arabic Lemma	عُلِم	عَلَّم	عَلَم	عَلِم	عِلم
Transli-Teration	[Eulim][4]	[Eal~am]	[Ealam]	[Ealim]	[Eilom]
English Meaning	be known/ be found out	teach/ instruct	flag/ banner/ badge	know/ find out	knowledge/ science

Less abstract than root but fully vocalized, each lemma defines a dictionary entry. Particularly, the verbs are reduced to the third masculine singular form of past tense. Each noun and verb are derived from a non-vocalized root according to a specific Arabic pattern. However, Arabic dictionaries are not sorted by lemmas but according to roots ordering.

3.1.1. Vocalization

Arabic words are vocalized with diacritics but only didactic documents or Koranic texts include full vocalization. This fact accentuates the ambiguity of

[2]http://www.qamus.org/morphology.htm
[3]http://siraj.sakhr.com/

[4]Herein, we opt to put transliteration between brackets [..].

words and requires paying more attention to the morphology and the word context. Table 2 gives a simple example for a token composed of only three consonants. One can consider different segmentations when removing diacritics.

Table 2. Four possible solutions[5] for the token [bsm] بسم.

Solution	Morphology	Vocalization	English meaning
1	Noun	[basom] بَسْم	Smiling
2	Verb	[basam] بَسَم	Smile
3	Prep + noun	[bi]بـ + [somi] سْم	In/by (the) name of
4	Prep + noun	[bi]بـ + [sam~] سَمّ	By/with poison

This aspect has been handled in a tagging system for classifying words in a non-vocalized Arabic text. The authors proposed to combine both morphological and syntax analyzer for part-of-speech annotation [1].

3.1.2. Agglutination

In Arabic text, a lexical unit is not easily identifiable from a graphic unit (word delimited by space characters or punctuation marks), but it frequently embodies a sequence of linked words. In addition to the core (stem), a word can be extended by attaching four kinds of affixes (antefixes, prefixes, suffixes and postfixes).

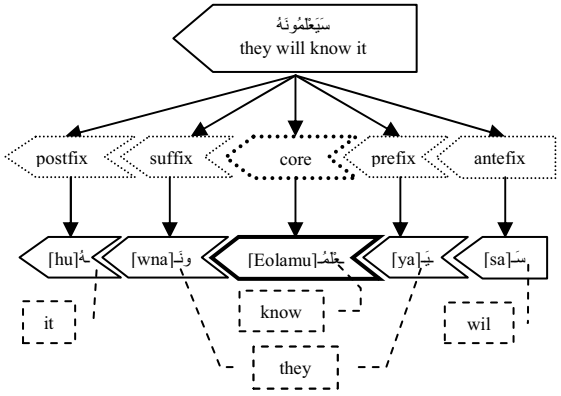

Figure 1. Segmentation of the Arabic agglutinated form [sayaEolamuwnahu] سَيَعْلَمُونَهُ.

In Figure 1, various kinds of affixes are attached to the core [Elm] علم in the agglutinated form of [sayaEolamuwnahu] سَيَعْلَمُونَهُ. This situation can make high ambiguity to extract the right core from an agglutinated form. In non-vocalized texts, morphology analysis become more difficult as illustrated above in Table 2.

3.2. Stemming Methods

According to the used approaches, two classes of Arabic stemming methods can be identified:

1. Light stemmers by removing the most common affixes.
2. Morphological analyzers by extracting each core according to a scheme.

3.2.1. Light Stemmers

Light stemming refers to the technique which truncates, from a word, a reduced list of affixes without trying to find roots. Effectiveness of this approach depends on the content of prefixes and suffixes lists. When, in English, one tries to find a stem by, mainly, removing conjugation suffixes, we have to deal in Arabic texts with ambiguous agglutinated forms that imply several morphological derivations. An analysis of such an approach can be found in [14]. This kind of stemmers can effectively deal with most practical cases, but in some ones we risk losing the right word. As an example, in the word [wafiy] وَفِي, one can read two agglutinated prepositions that mean to "and in" but another will consider a noun with meaning to "faithful/complete".

3.2.2. Morphological Analyzers

In morphological analysis, we try to extract more complete forms according to vocalization variation and derivation patterns knowledge. We can distinguish two categories of analyzers according to the nature of desired output unit:

1. Root-based stemmers.
2. Lemma-based stemmers.

The choice between the two stemmers depends on how further stemming results, in IR tasks or in language modeling, will be used.

In first category, the Khoja stemmer[6], which attempts to find root for an Arabic word, has been proposed in [13]. A list of roots and patterns is used to determine the right form. This approach produces abstract roots which reduce significantly the dimension of documents features space. However, it conducts to a confusion of divergent meaning embedded in a unique root. For example, stemming of the word [sayaEolamuwnahu] سَيَعْلَمُونَهُ cited above in Figure 1 must deduce the root-stem [Elm] علم with possible meaning to "to know or science" but never to "flag", the third sense in Table 1.

In second category, Buckwalter [4] has developed a set of Arabic lexicons[7] with rules for legal combinations of lemma-stems and affixes forms. P. Brihaye has developed a java package[8] based on the Arabic morphological analyzer of Buckwalter, where several stemming solutions are proposed for each word. From this analyzer, one can develop, under some

[5] According to the Buckwalter Arabic morphological analyzer (ver 1.0 2002).

[6] http://zeus.cs.pacificu.edu/shereen/ArabicStemmerCode.zip

[7] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49

[8] http://www.nongnu.org/aramorph/english/index.html

considerations, a lemma-based stemmer. This approach will be detailed and tested hereafter.

3.3. Lemma-Based Stemmer

We propose an algorithm for lemma-based stemming that is called the Brahmi-Buckwalter Stemmer and referred henceforth as *BBw*. Based on the resources of the Buckwalter morphological analyzer, two main contributions can be reported for the *BBw* stemmer:

1. Normalization preprocessing.
2. Stem selection with morphological analysis.

3.3.1. Pre-Processing

This step is performed for normalizing the input text. Then the obtained list of tokens will be processed by the Buckwalter morphological analyzer.

- Convert to UTF-8 encoding.
- Tokenize text respecting the standard punctuation.
- Remove diacritics and tatweel (ـ).
- Replace initial alef with hamza above or below (أ or إ) with bar-alef (ا).
- Replace final waw or yeh with hamza above (ؤ or ئ) with hamza (ء).
- Replace maddah (آ) or alef-waslah (ٱ) with bar-alef (ا).
- Replace two bar-alef (اا) with alef-maddah (آ).
- Replace final teh marbuta (ة) with heh (ه).
- Remove final yeh (ي) when the remaining stem is valid.
- Remove non-Arabic letters and stop-words.

3.3.2. Stem Selection

When an input token (in-token) is processed by the Buckwalter morphological analyzer, three cases can be reported: 1). A unique solution is given to the in-token according to a specific pattern. 2). Multiple solutions are found corresponding to several patterns and lexicons entries. 3). No solution can be attributed to the in-token. The actions, that the *BBw* stemmer must undertake, will be detailed bellow:

1. *Unique Solution:* The *BBw* stemmer retains only the non-vocalized lemma of the solution. Each solution without stem (i.e., contains only affixes) is ignored and therefore the in-token is considered as a stop-word.
2. *Multiple Solutions:* The *BBw* stemmer treats all the proposed solutions as a set of separated unique solutions and thus retains all non-vocalized lemmas. Note that eliminating vocalization may unify some stems and so reduces the number of stem-solutions. For example, Table 2 gives four vocalized solutions for the token [bsm] بسم but when removing diacritics, the BBw stemmer will identify only two confused stems {[bsm] بسم, [sm] سم}.
3. *No Solution:* When no solution is found for the in-token, different reasons can be raised:
   a. The in-token is wrong and it did not imply any Arabic lemma.
   b. The in-token corresponds to a proper name (person, city, etc.,) that has no entry in the dictionary.
   c. The in-token is a correct Arabic word but it is not yet included in the current release of Buckwalter morphological analyzer.

We opt herein to keep the unknown in-token as a stem solution and so a new vocabulary entry will be added. This will allows us to analyze our algorithm, with the basic lexicon of Buckwalter analyzer, on real world corpora.

3.3.3. Confusion Degree Measure

When the morphological analysis of the in-token implies multiple solutions, *BBw* stemmer produces multiple lemma-stems. For an in-token (t_i) the stemmer will give (d_i) distinct stems. Then, we define, by $d_i \in \{1,2,3,\ldots\}$, the degree of the (t_i)-solutions multiplicity. The total number of tokens in a non-empty text collection (S) is denoted by $L \neq 0$. So, the confusion degree (C), in the collection (S) stemmed by the algorithm (R), is defined as:

$$C(S|R) = \frac{1}{L}\sum_{i=1}^{L} d_i \tag{1}$$

for example, $C(S|Khoja)=1$, since the Khoja stemmer gives at most one stem for each token from any dataset S. This is an ideal situation for a stemming process but when applying the *BBw* stemmer, the confusion degree C will be increased. We proposed this measure $C(S|R)$ for assessing the lexical ambiguity in Arabic texts. For a human Arabic reader, this problem will easily be solved with a semantic consideration guided by the context.

With *BBw* stemming, note that all possible stems are equitably related to their in-token. At this stage, we have no additional knowledge to select the good stem. We think that the relevant solution can be weighted later by a co-occurrence computation in large corpora as a local context. This will be feasible with LDA topic modeling.

4. Topic Models

LDA is a generative topic model for text documents [2]. Based on the classical "bag of words" assumption, a topic model considers each document as a mixture of topics where a topic is defined by a probability distribution over words. The distribution over words within a document is given by:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j) \tag{2}$$

where $P(w|z)$ defines the probability distribution over words w given topic z. $P(z)$ refers to the distribution over topics z in a collection of words (document). The probability density of a T (number of topics) dimensional dirichlet distribution over the multinomial distribution $p=(p_1, \ldots, p_T)$ is defined by:

$$Dir(\alpha_1,\ldots,\alpha_T) = \frac{\Gamma\left(\sum_j \alpha_j\right)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} p_j^{\alpha_j - 1} \quad (3)$$

The parameters of such a distribution are specified by $\alpha_1 \ldots \alpha_T$ where each hyper-parameter α_j can be interpreted as a prior observation count for the number of times topic j is sampled in a document, before having observed any actual word from that document. The theoretical basis of LDA model are mainly taken from [2, 21] where more details and interpretations can be found.

For a given number of topics T, LDA model is learned from a collection of documents defined as follows:

N : number of words in vocabulary.
M : number of document in corpus.
T : number of topics, given as input value.

In addition, we define:
$P(z)$: distribution over topics z in a particular document.
$P(w/z)$: probability distribution over words w given topic z.

Then, we can define a generative process as follows:

For each document d=1 to M (in dataset) do:
1. *Sample mixing probability* $\theta_d \sim Dir(\alpha)$
2. *For each word* $w_{di} = 1$ *to N (in vocabulary) do:*
   2.a. Choose a topic $z_{di} \in \{1, .., T\} \sim Multinomial(\theta_d)$
   2.b. Choose a word $w_d \in \{1, .., N\} \sim Multinomial(\beta_{zdi})$

where α is a Dirichlet symmetric parameter and $\{\beta_i\}$ are multinomial topic parameter. Each β_i assigns a high probability to a specific set of words that are semantically consistent. This distribution over vocabulary is referred to as topic. In the present work, we use the LDA implementation of LingPipe[9] package which is based on Gibbs sampling algorithm for parameters estimation.

The number of topics affects directly the significance of the LDA model training results. Since the number of topics (T) is given as an input parameter for training the LDA model, several methods were proposed to select a suitable T [2]. An evident approach is to choose T that lead to best generalization performance in the followed tasks.

5. Arabic Datasets

Due to the unavailability of free Arabic resources, we have opted to build our own experimentation datasets. For this aim, we developed a Web-Crawler[10] application to collect newspaper articles from several Arabic websites.

5.1. Datasets Description

In this study, we present three real-world corpora based on Echorouk[11], Reuters[12] and Xinhua[13] Web-articles which relating to the 2007-2009 period.

Table 3. Description of three datasets relating to Echorouk, Reuters and Xinhua Web-articles.

Feature \ Dataset	Ech-11k	Rtr-41k	Xnh-36k
# Articles	11,313	41,251	36,696
# Characters	48,247,774	111,478,849	85,696,969
# Tokens	4,388,426	10,093,707	7,532,955
# Arabic Tokens	3,341,465	7,892,348	6,097,652
Average # Tokens/Article	387.9	244.7	205.3
# Categories	8	6	8

Table 4. Distribution of the three datasets over categories.

	Category \ Dataset	Ech-11k (Echorouk)	Rtr-41k (Reuters)	Xnh-36k (Xinhua)
1	**World**	2,274	10,000	9,465
2	**Economy**	816	10,000	6,862
3	**Sport**	3,554	10,000	1,132
4	**Middle-East**	-	10,000	9,822
5	**Science-Health**	-	889	1,993
6	**Culture-Education**	566	-	1,508
7	**Algeria**	2,722	-	-
8	**Society**	808	-	-
9	**Art**	315	-	-
10	**Religion**	258	-	-
11	**Entertainment**	-	362	-
12	**China**	-	-	4,654
13	**Tourism-Ecology**	-	-	1,260
Total		11,313	41,251	36,696

Each article is saved with UTF-8 encoding in a separated text file where the first line is reserved to its title. A brief description about the collected datasets is given in Table 3. Each dataset is labeled according to the editor categorization. Table 4 describes the datasets distribution over published categories.

5.2. Arabic Stemming

As described in previous sections, the Khoja root-based stemmer and the *BBw* lemma-based stemmer were applied on the three datasets described above. Relating to each corpus stemming, different vocabulary sizes are produced. As common ascertainments for the three datasets, Table 5 shows

[9]Available at http://alias-i.com/lingpipe/index.html

[10]Developed with Amine Roukh and Abdelkadir Sadoki in Mostaganem University.

[11]Algerian newspaper, http://www.echoroukonline.com/

[12]International news agency, http://ara.reuters.com/

[13]Chinese news agency, http://arabic.news.cn/

that with lemma-based stemming, an important enhancement is reported for the vocabulary size. It is an expected result since *BBw* algorithm retains the unrecognized in-token and provides less abstract stems.

Table 5. Vocabulary sizes related to two stemming approaches.

Stemmer \ Dataset	Ech-11k	Rtr-41k	Xnh-36k
Khoja (root-based)	22,543	38,773	34,304
BBw (lemma-based)	48,789	62,843	63,524

For assessing performance of *BBw* stemmer when dealing with ambiguous forms, we report in Table 6 the confusion degree as defined in equation 1.

Table 6. Analysis of solution multiplicity with BBw stemming applied in three datasets.

Dataset	Ech-11k		Rtr-41k		Xnh-36k	
Solution Multiplicity	**# Token**	**%**	**# Token**	**%**	**# Token**	**%**
1	3,021,314	90.42	7,092,343	89.86	5,532,178	90.73
2	287,168	8.59	694,572	8.80	495,599	8.13
3	30,615	0.92	95,202	1.21	64,117	1.05
4	2,174	0.07	9,179	0.12	5,612	0.09
5	194	0.01	1,051	0.01	146	0.00
6	0	0.00	1	0.00	0	0.00
Confusion Degree	1.106		1.116		1.105	

Table 6 shows that close to 90% of analyzed tokens provide unique solutions and that the maximum confusion degree, in the three real-world corpora, has not exceeded 1.12 (last line in Table 6). This indicates that when preserving all word senses by reporting all multiple solutions, the *BBw* stemmer did not cause lexical ambiguity.

6. Experiments

For the validation of our study, the three datasets (Echorouk, Reuters and Xinhua) are firstly stemmed. Second, latent topics are learned with Lingpipe LDA implementation for different topic numbers. Then, a supervised classification is performed in the reduced documents distribution over topics. It is worth pointing out that: 1). For each dataset, two stemmers (Khoja and *BBw*) are applied, 2). In topic modeling, all datasets are used as unlabeled collections, and 3). For supervised classification, 3/4 of each labeled dataset is used for the training process and the remaining documents as the test-set.

We apply the SVM algorithm [23] with LIBSVM[14] package for documents classification. Two measures were used for performance evaluation: On the one hand, the well-known recognition rate which is computed by dividing the number of the well-classified articles by the test-collection size. On the other hand, we define a ratio (rSV) of the number of Support Vectors (SV) in the total number of training examples. A lower rSV ratio corresponds to a better generalization performance [5]. In addition, the evaluation is completed by applying SVM classification in the full words space, with the Term Frequency (TF) measure. As SVM training parameter, the simple linear kernel is used with a cost parameter fixed to 10.

To appreciate the advantage of latent topics for discovering semantic in texts we propose to compute the categories distribution over learned topics. A confusion matrix (category\topic) can be obtained by adding document distributions in the same category.

7. Results and Discussion

7.1. Classification in Topics Space

The results in Table 7 show that the best classification performances are obtained when training LDA for 64 or 100 topics. Although the recognition rate measure (Reco.) does not give a significant difference between Khoja and *BBw* stemmers, but when considering generalization performance with *rSV* measure, two preliminary observations can be raised: Firstly, topic modeling is more efficient than classification in words space. Secondly, the *BBw* lemma-stemmer improves classification when comparing to Khoja root-stemmer.

Table 7. Evaluation of SVM classification on the three datasets.

Dataset	Stem.	Perform.	8	16	32	64	100	TF
Ech-11k	**Khoja**	**Reco.**	75.2	80.7	82.1	82.2	82.2	78.6
		rSV	46.5	41.0	35.7	35.8	36.6	41.4
	BBw	**Reco.**	80.5	81.7	83.6	**83.7**	83.7	81.8
		rSV	37.3	38.3	33.1	**32.7**	33.5	44.3
Rtr-41k	**Khoja**	**Reco.**	78.9	84.6	84.8	85.6	88.2	83.0
		rSV	38.5	27.6	26.3	24.1	23.4	25.2
	BBw	**Reco.**	78.9	82.1	82.6	88.3	**89.8**	84.5
		rSV	36.9	33.3	29.1	21.7	**20.8**	25.4
Xnh-36k	**Khoja**	**Reco.**	58.9	67.9	69.4	74.7	75.5	70.0
		rSV	53.3	48.8	45.8	39.4	38.0	37.9
	BBw	**Reco.**	59.0	66.5	72.0	74.9	**76.7**	73.5
		rSV	54.6	49.7	41.7	37.4	**35.9**	38.8

7.2. Finding Topics in Newspaper Articles

We illustrate some results of LDA modeling applied on the three datasets. Text preprocessing is performed by the *BBw* lemma-stemmer. Table 8 gives the titles of 16 learned topics[15]. The latent topics are titled with human assessment according to its relevant terms.

Table 9 shows the reuters categories distributions over 16 main topics. For example, one can easily discover that the main subjects in sport category of reuters articles during 2007-2009 were football and beijing olympic games.

[14] LIBSVM-2.89 is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm+zip

[15] More details are available at https://sites.google.com/site/abderrezakbrahmi/

Table 8. Titles of the 16 main latent topics in the three datasets.

Topic	Ech-11k	Rtr-41k	Xnh-36k
1	Algerian economy	Financial markets	Culture
2	Judicial affairs	Palestine	USA and UN
3	Local Algeria	Health	Iraq
4	Algerian international affairs	Football	Iranian nuclear
5	Security affairs	Iran	Economy
6	Algerian football	Business	Anti-terrorism affairs
7	Iraq and Iran	Somalia and Sudan	Palestine
8	Palestine	Beijing Olympic games	China politics
9	Algerian football team	Petrol and gas	Football
10	Art and culture	Judicial affairs	Chinese economy
11	General terms	European football	Markets
12	Health	Military actions in Iraq and Afghanistan	Arabic region
13	Education	Tennis	Seism in China
14	Sport organization	Iranian nuclear	Transport and travelling
15	Islam	Middle-east	Health
16	Government projects	USA politics	Chinese coope-ration

Table 9. Distribution of reuters categories over 16 latent topics.

Category \ Topic	Middle-East	Business	World	Sport	Enter-Tainment	Science
Financial markets	0.6	**30.8**	1.1	0.5	1.7	2.0
Palestine	**18.1**	0.6	0.8	0.4	3.5	0.5
Health	7.0	2.3	**11.6**	1.7	**30.4**	**50.0**
Football	0.3	0.3	0.3	**25.0**	1.0	0.5
Iran	3.2	0.7	8.9	0.6	1.9	1.1
Business	2.0	**29.4**	1.2	1.0	4.4	2.1
Somalia and Sudan	**13.5**	1.0	5.5	0.5	1.4	2.0
Beijing Olympic games	0.4	1.1	2.2	**10.7**	**13.3**	2.8
Petrol and gas	1.8	**18.0**	1.7	0.4	1.7	5.8
Judicial affairs	**11.5**	1.6	**10.5**	1.5	**19.6**	3.0
European football	0.6	0.6	2.2	**23.4**	2.5	1.1
Military actions in Iraq and Afghanistan	**12.4**	0.7	**15.6**	0.5	2.3	1.4
Tennis	0.2	0.4	0.4	**29.8**	0.7	0.4
Iranian nuclear	3.7	1.7	**10.8**	0.6	1.1	2.1
Lebanon and Syria	**11.9**	2.1	9.4	0.8	5.1	1.2
USA politics	**12.7**	8.8	**17.7**	2.5	9.5	**24.0**

In addition, words can be analyzed by finding their different contexts when increasing the number of topics for LDA training. In Table 10, an analysis example of the word [slAm] سلام is given when modeling the three datasets according to 100 topics. Two senses can be assigned to [slAm] سلام, either peace or greeting /salutation.

Table 10. The topics related to ([slAm] سلام) in three dataset.

Ech-11k	Rtr-41k	Xnh-36k
Arab league, Palestinian dialogue.	Somali actors, Nobel price, Middle-East negotiations, Sudan events.	UN missions, Jordan's peace effort, Middle-East negotiations, Coal mines in Australia, Sudan events.

It is clear, in Table 10, that the different contexts related to this word implies the first sense except for the forth topic in Xnh-36k corpus. In fact, the topic *Coal* mines in Australia is connected to another word, [slAmh] سلامه, which has two stemming solutions ([slAm] سلام and [slAmh] سلامه). The second solution, that means safety, is highly correlated to security and safety requirements in coal mines.

7.3. Summary

For Arabic topic discovering, LDA model was applied with Gibbs sampling algorithm and three real-world corpora have been collected from Arabic newspaper articles (Echorouk, Reuters and Xinhua).

Experiments show that performances, in terms of accuracy and generalization, of SVM-classification in the reduced topics space outperform classification in full words space. Furthermore, the proposed stemming approach proves its efficiency in both Arabic topic modeling and supervised classification. In fact, root-based stemming, such as Khoja algorithm, gives good performances in some Arabic IR tasks. However, trying to characterize Arabic text with abstract roots for topics learning, may affect model stability [8, 12]. In opposite to *BBw*, root-based stemmers provide, for one in-token, an abstract stem and so, several senses may be merged in a unique index entry.

It is worth pointing out that it is difficult to understand how one can assess semantic aspects in Arabic texts without sufficient linguistic knowledge or any help from an Arabic expert [14, 15]. This study shows that effective advances in Arabic IR and topic modeling must be developed with a close collaboration between computer scientists and Arabic language experts.

8. Conclusions and Future Work

Arabic language has a highly inflected morphology with two particular characteristics: agglutination and non-vocalization.

In addition to the adaptation of LDA to Arabic topic modeling, two main contributions were described in this paper: Firstly, we have introduced a lemma-based stemmer with specific normalization and morphological analysis. The *BBw* algorithm was developed and analyzed as a linguistic preprocessing for topic modeling and supervised classification. A confusion degree measure is proposed to assess the effect of reporting all possible solutions for one in-token. When applying this measure in different Arabic datasets, we have shown that our stemming approach may preserve the semantic embedded in Arabic texts without compromising lexical characterization.

Secondly, three real-world corpora were tested for Arabic topic modeling and documents categorization. Tens of thousands of Web-articles were automatically crawled from Echorouk, Reuters and Xinhua. The

variety of writing styles allows us to validate the proposed lemma-based stemmer.

As future works, we plan to improve *BBw* stemmer performance by handling some irregular forms and enhancing lexicons with proper names (person, location and organization). Further efforts must be oriented to topics visualization and topic-based application for end-user retrieval system.

Acknowledgments

We gratefully acknowledge Professor Patrick Gallinari, director of LIP6 laboratory (Paris-France), for his valuable advice to realize this work.

References

[1] Al-Taani A. and Abu Al-Rub S., "A Rule-Based Approach for Tagging Non-Vocalized Arabic Words," *The International Arab Journal of Information Technology*, vol. 6, no. 3, pp. 320-328, 2009.

[2] Blei D., Ng A., and Jordan M., "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 993-1022, 2003.

[3] Brants T., Chen F., and Farahat A., "Arabic Document Topic Analysis," *in Proceedings of Workshop on Arabic Language Resources and Evaluation*, Spain, pp. 1-4, 2002.

[4] Buckwalter T., "Buckwalter Arabic Morphological Analyzer Version 1.0," available at: http://www.ldc.upenn.edu/Catalog/docs/LDC2002L49, last visited 2002.

[5] Burges C., "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[6] Chemudugunta C., Smyth P., and Steyvers M., "Combining Concept Hierarchies and Statistical Topic Models," *in Proceedings of the 17th ACM Conference on Information and Knowledge Management*, USA, pp. 1469-1470, 2008.

[7] El-Halees A., "Arabic Text Classification Using Maximum Entropy," *The Islamic University Journal*, vol. 15, no. 1, pp. 157-167, 2007.

[8] El-Kourdi M., Bensaid A., and Rachidi T., "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," *in Proceedings of COLING 20th Workshop on Computational Approaches to Arabic Script-Based Languages*, pp. 51-58, 2004.

[9] Hofmann T., "Probabilistic Latent Semantic Analysis," *in Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pp. 289-296, 1999.

[10] Iwata T., Yamada T., and Ueda N., "Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents," *in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, USA, pp. 363-371, 2008.

[11] Jagaralamudi J. and Daumé H., "Extracting Multilingual Topics from Unaligned Corpora," *in Proceedings of the European Conference on Information Retrieval*, UK, pp. 444-456, 2010.

[12] Kadri Y. and Nie J., "Effective Stemming for Arabic Information Retrieval," *The Challenge of Arabic for NLP/MT, International Conference at the British Computer Society*, pp. 68-74, 2006.

[13] Khoja S. and Garside R., "Stemming Arabic Text," *Computing Department,* Lancaster University, Lancaster, 1999.

[14] Larkey L., Ballesteros L., and Connell M., "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis," *in Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, USA, pp. 275-282, 2002.

[15] Larkey L., Feng F., Connell M., and Lavrenko V., "Language-Specific Models in Multilingual Topic Tracking," *in Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, USA, pp. 402-409, 2004.

[16] Moukdad H., "Stemming and Root-Based Approaches to the Retrieval of Arabic Documents on the Web," *Webology*, vol. 3, no. 1, article 22, 2006.

[17] Nallapati R. and Cohen W., "Link-pLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs," *in Proceedings of the International Conference on Weblogs and Social Media*, USA, pp. 84-92, 2008.

[18] Newman D., Hagedorn K., Chemudugunta C., and Smyth P., "Subject Metadata Enrichment Using Statistical Topic Models," *in Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, USA, pp. 366-375, 2007.

[19] Oard D. and Gey F., "The TREC-2002 Arabic/English CLIR Track," *in Proceedings of the 11th Text REtrieval Conference, National Institute of Standards and Technology*, pp. 81-93, 2002.

[20] Porteous I., Newman D., Ihler A., Asuncion A., Smyth P., and Welling M., "Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation," *in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, USA, pp. 569-577, 2008.

[21] Steyvers M. and Griffiths T., "Probabilistic Topic Models," *in Proceedings of Latent Semantic Analysis: A Road to Meaning*, Laurence Erlbaum Associates, pp. 1-15, 2007.

[22] Tuerlinckx L., "La Lemmatisation de L'arabe Non Classique," *in Proceedings of the* 7es *Journées Internationales d'Analyse Statistique des Données Textuelles*, pp. 1069-1078, 2004.
[23] Vapnik V., *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.
[24] Zhang D., Mei Q., and Zhai C., "Cross-Lingual Latent Topic Extraction," *in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1128-1137, 2010.

Abderrezak Brahmi received his engineering degree in 1994 from the Oran University, Algeria, and his MSc degree 2005 from University of Sciences and Technology of Oran-USTO. He is actually a lecturer at the University of Mostaganem, Algeria, and prepares a PhD in intelligent Web retrieval. His research interest includes semantic Web, natural language processing and machine learning.

Ahmed Ech-Cherif holds both a MS and PhD degrees from Rensselaer Polytechinc Institue New York, USA. After graduation, he was employed by Gerneral Motors Research Labs, USA as a senior research scientist. He has published several journal publications in machine learning, data mining and mathematical programming. Currently, he is an associate professor of computer science at the University of Sciences and technology, Algeria.

Abdelkader Benyettou received his engineering degree in 1982 from the Institute of Telecommunications of Oran and the MSc degree in 1986 from the University of Sciences and Technology, Algeria. In 1987, he joined the Computer Sciences Research Center of Nancy, France, where he worked until 1991 on Arabic speech recognition by expert systems, and received his PhD in electrical engineering in 1993 from the USTO University. He is actually a professor at USTO University since 2003. Currently, he is a researcher director of the Signal-Image-Speech-Laboratory, USTO.